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

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

Affiliations:

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

[2] Agora Cancer Research Centre, Lausanne, Switzerland

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

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

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

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

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

* Corresponding author: david.gfeller@unil.ch

Abstract

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

cancer cohort, EPIC-ATAC accurately infers the immune contexture of the main breast cancer subtypes.

Introduction

Gene regulation is a dynamic process largely determined by the physical access of chromatin-binding factors such as transcription factors (TFs) to regulatory regions of the DNA (*e.g.,* enhancers and promoters) (1). 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 (1,2). Over the last decade, Assay for Transposase-Accessible Chromatin (ATAC-Seq) (3) has become a reference epigenomic technique to profile chromatin accessibility and the activity of gene regulatory elements in diverse biological contexts including cancer (4) and across large cohorts (5). 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 (6,7), as well as the formalin-fixed paraffin-embedded (FFPE)-ATAC protocol adapted to FFPE samples. The reasonable cost and technical advantages of these protocols foreshadow an increased usage of ATAC-Seq in cancer studies.

Most biological tissues are composed of multiple cell types. For instance, tumors are complex ecosystems including malignant and stromal cells as well as a large diversity of immune cells. This cellular heterogeneity, in particular the presence of specific immune cell types, impacts tumor progression as well as response to immunotherapy (8–10). Most existing ATAC-Seq data from tumors were performed on bulk samples, thereby including information from both cancer and non-malignant cells. Precisely quantifying the proportions of different cell types in such samples represents therefore a promising way to explore the immune contexture and the composition of the tumor micro-environment (TME) across large cohorts. Carefully assessing cell-type heterogeneity is also important to handle confounding factors in genomic analyses in which samples with different

51 cellular compositions are compared. Recently, single-cell ATAC-Seq (scATAC-Seq) has been developed
52 to explore cellular heterogeneity with high resolution in complex biological systems (11–13).
53 However, the resulting data are sensitive to technical noise and such experiments require important
54 resources, which so far limits the use of scATAC-Seq in contrast to bulk sequencing in the context of
55 large cohorts.
56
57 In the past decade, computational deconvolution tools have been developed to predict the
58 proportion of diverse cell types from bulk genomic data obtained from tumor samples (14–22). A
59 large number of these tools model bulk data as a mixture of reference profiles identified in purified
60 cell populations for each cell type. The accuracy of the predictions of cell-type proportions relies on
61 the quality of these reference profiles as well as on the use of cell-type specific markers (14). A
62 limitation of most deconvolution algorithms is that they do not predict the proportion of cell types
63 that are not present in the reference profiles (here referred to as 'uncharacterized' cells). In the
64 context of cancer samples, these uncharacterized cell populations include malignant cells whose
65 molecular profiles differ not only from one cancer type to another, but also from one patient to
66 another even within the same tumor type (5). A few tools consider uncharacterized cells in their
67 deconvolution framework by using cell-type specific markers not expressed in the uncharacterized
68 cells (17,21,23,24). These tools include EPIC (Estimating the Proportion of Immune and Cancer cells)
69 which simultaneously quantifies immune, stromal, vascular as well as uncharacterized cells from bulk
70 tumor samples (17,25).
71
72 Most deconvolution algorithms have been initially developed for transcriptomic data (RNA-Seq data)
73 (17–19,21,26–29). More recently they have been adapted for other omics layers such as methylation
74 (30–33) and proteomics (34) or chromatin accessibility. For the latter, a specific framework called
75 DeconPeaker (20) was developed to estimate cell-type proportions from bulk samples.
76 Deconvolution tools developed initially for other omics modalities, such as RNA-Seq, can also be

applied on ATAC-Seq if appropriate ATAC-Seq profiles are provided to the tool. For example, the popular deconvolution tool, CIBERSORT (26), was used to deconvolve leukemic ATAC-Seq samples (35). Other methods have been proposed to decompose ATAC-Seq bulk profiles into subpopulation-specific profiles (36,37) or compartments (38). However, these methods have more requisites: (i) the integration of the ATAC-Seq data with single-cell or bulk RNA-Seq (36,37) and HIChIP data (36) or, (ii) subsequent feature annotation to associate compartments with cell types or biological processes (38).

The application of existing bulk ATAC-Seq data deconvolution tools to solid tumors is limited. First, current computational frameworks do not quantify populations of uncharacterized cell types. Second, ATAC-Seq based markers (*i.e.,* chromatin accessible regions called peaks) and reference profiles generated so far have been derived in the context of hematopoietic cell mixtures (20,35). Markers and profiles for major populations of the TME (*e.g.,* stromal and vascular cells) are thus missing. While cell-type specific markers have been identified from scATAC-Seq data (2), not all TME-relevant cell types are covered (*e.g.*, lack of scATAC-Seq data from neutrophils due to extracellular traps formation). Also, these markers have not been curated to fulfill the requirements of tools such as EPIC to quantify uncharacterized cells (*i.e.*, markers of a cell-type should not be accessible in other human tissues).

In this study, we collected ATAC-Seq data from pure cell types to identify cell-type specific marker peaks and to build reference profiles from most major non-malignant cell types typically observed in tumors. These data were integrated in the EPIC (17) framework to perform bulk ATAC-Seq samples deconvolution (Figure 1). Applied on peripheral blood mononuclear cells (PBMCs) and tumor samples, the EPIC-ATAC framework showed accurate predictions of the proportions of non-malignant and malignant cells with similar or higher performances than other existing tools.

103 ## Results

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

106 A key determinant for accurate predictions of cell-type proportions by most deconvolution tools is
107 the availability of reliable cell-type specific markers and reference profiles. To identify robust
108 chromatin accessibility marker peaks of cancer relevant cell types, we collected 564 samples of
109 sorted cell populations from twelve studies including eight immune cell types (B cells (35,39,40),
110 CD4+ T cells (35,40–43), CD8+ T cells (35,39–41,43), natural killer (NK) cells (35,39), dendritic cells
111 (DCs) (39,41,44), macrophages (40,41), monocytes (35,39,40,44,45) and neutrophils (46,47), as well
112 as fibroblasts (41,48) and endothelial (41,49) cells (Figure 1 box 1, Figure 2A, Supplementary Table 1).
113 To limit batch effects, the collected samples were homogeneously processed from read alignment to
114 peak calling. For each cell type, we derived a set of stable peaks, *i.e.,* peaks observed across samples
115 and studies (see Materials and Methods).

116 These peaks were then used to perform pairwise differential analysis to identify marker peaks for
117 each cell type (Figure 1, box 2). To ensure that the cell-type specific marker peaks are not accessible
118 in other human tissues, we included in the differential analysis ATAC-Seq samples from diverse
119 human tissues from the ENCODE data (50,51) (Supplementary Figure 1). To select a sufficient number
120 of peaks prior to peak filtering, the top 200 peaks recurrently differentially accessible across all cell-
121 type pairs were selected as cell-type specific markers (see Materials and Methods). Using the human
122 atlas study (2), markers with potential residual accessibility in human tissues were then filtered out
123 (Figure 1, box 3, see Materials and Methods). The resulting marker peaks specific to the immune cell
124 types were considered for the deconvolution of PBMC samples (PBMC markers). For tumor bulk
125 sample deconvolution, the list of markers was further refined based on the correlation patterns of
126 the markers in tumor bulk samples from diverse cancer types from The Cancer Genome Atlas (TCGA)
127 (5) (Figure 1, box 4, see the Material and methods). The latter filtering ensures the relevance of the
128 markers in the TME context since cell-type specific TME markers are expected to be correlated in

129 tumor bulk ATAC-Seq measurements (52). 716 markers of immune, fibroblasts and endothelial cell
130 types remained after the later filtering and were considered for the deconvolution of bulk tumor
131 samples (TME markers).
132 To assess the quality and reproducibility of these markers, we performed principal component
133 analysis (PCA) based on each set of marker peaks. Computing silhouette coefficients based on the
134 cell-type classification and on the study of origin showed that samples clustered by cell type and not
135 by study of origin (averaged silhouette coefficients above 0.45 for cell type and around 0 for study of
136 origin). Two-dimensional UMAP representations of the samples confirmed this observation (Figure
137 2B). These results indicate limited remaining batch effects after data processing and marker
138 selection.
139 We then used the collected samples to generate chromatin accessibility profiles by computing the
140 average of the normalized counts for each peak in each cell type as well as peak variability in each
141 cell type (17) (see Material and methods). Figure 2C represents the average chromatin accessibility of
142 each marker peak in each cell type of the reference dataset and highlights, as expected, the cell-type
143 specificity of the selected markers (see also Supplementary Tables 2 and 3), which was confirmed in
144 independent ATAC-Seq data from sorted cells and single-cell ATAC-Seq samples from blood and
145 diverse human tissues (Figure 2D and 2E, see Materials and methods).
146

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

148 To characterize the different marker peaks, we annotated them using ChiPSeeker (53). We observed
149 that most of the markers are in distal and intergenic regions (Figure 2F), which is expected
150 considering the large proportion of distal regions in the human genome and the fact that such
151 regions have been previously described as highly cell-type specific (35). We also noticed that 7% of
152 the PBMC and TME marker peaks are in promoter regions in contrast to 4% when considering
153 matched genomic regions randomly selected in the set of peaks identified prior to the differential

154 analysis (see Material and methods), which suggest enrichment in our marker peaks for important
155 regulatory regions.
156 To assess the biological relevance of the marker peaks, we associated each marker peak to its
157 nearest gene using ChIP-Enrich based on the "nearest transcription start site (TSS)" locus definition
158 (54) (Supplementary Tables 4 and 5). Nearest genes reported as known marker genes in public
159 databases of gene markers (*i.e.,* PanglaoDB (55) and CellMarker (56)) are listed in Table 1.
160 In each set of cell-type specific peaks, we observed an overrepresentation of chromatin binding
161 proteins (CBPs) reported in the JASPAR2022 database (57) (using Signac (58) and MonaLisa (59) for
162 assessing the overrepresentation) and the ReMap catalog (60) (using RemapEnrich, see Material and
163 Methods). Overrepresented CBPs also reported as known marker genes in the PanglaoDB and
164 CellMarker databases are listed in Table 1. Detailed peaks annotations are summarized in
165 Supplementary Tables 4 and 5.
166 Based on the "nearest TSS" annotation, we tested, using ChIP-Enrich (54), whether each set of cell-
167 type specific marker peaks was enriched for regions linked to specific biological pathways (GO
168 pathways). Figure 2G highlights a subset of the enriched pathways that are consistent with prior
169 knowledge on each cell type. Some of these pathways are known to be characteristic of immune
170 responses to inflammatory or tumoral environments. The complete list of enriched pathways is listed
171 in the Supplementary Tables 6 and 7. Overall, these analyses demonstrate that the proposed cell-
172 type specific marker peaks capture some of the known biological properties associated to each cell
173 type.

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

CD8_Tcells	MKI67 JAML MAML2 KLRD1 NELL2 LAG3 PPP1R13B PTPRC LYST CASP8 CD8A CD8B CD96 BTLA GZMA THEMIS ETV1	ETV1 FOXP3 TBX21 FOXP1 EOMES CREM IRF4 ZEB1 ARNT JUNB TCF7
NK	PRF1 ZBTB16 KLRD1 SPN CD226 SH2D1B CD247 IL2RB CXCR4 NMUR1 GNLY ZAP70 TXK	EOMES TBX21 NFIL3 FOS JUN
DCs	C12orf75 LYZ APP CD8A RIOX2 NFKB1 QDPR ABCG2 PRELID2 DST CD36 IDO2 PCMTD1	SPIB IRF8 MYB NR4A1 REL CUX2 FOXO1 ETV6 IRF5 BATF3 RUNX2
Neutrophils	TLE3 CA4 CYP4F3 CEACAM8 PGLYRP1 FPR1 CTSS ALPL PI3 MMP9 CXCR1 DRC1 ASPRV1 LTF MGAM SLC25A37	FOS
Monocytes	VENTX GLT1D1 CLEC4E CARS2 SLC24A4 C16orf74 FFAR2 STXBP2 NLRP3 CYRIA CMTM7 TGFBI DIAPH1 VCAN MCTP1 IFNGR1 STX11 CAPZA2 CD36 MTSS1 DENND3 ASAH1 TNFRSF10B BNIP3L NACC2 MAMDC2 FBP1	CEBPA CEBPD CEBPB CEBPE SPI1 VENTX JUND RXRA TCF7L2
Macrophages	CXCL12 PSAP P2RY6 SLCO2B1 CMKLR1 MMP19 LGMN CLEC10A C5AR1 FPR3 LILRB4 RGL1 SIGLEC1 MMP9 CD80	STAT1 SPI1 FOSL2 FOS SPIC
Endothelial	FAM107B ROBO4 FLI1 ACVRL1 FLT1 DOCK9 ABCC1 S1PR1 ELOVL1 PLPP3 ASAP2 SNRK ECSCR ARAP3 LAMA4 BMP6 SERPINE1 LAMB1 DOCK4 NOS3	ETV2 ELF1 FLI1 ELK3 FOSB ETS1 ERG GATA2 ZEB1 ETS2 FOXC1 SOX18
Fibroblasts	LOX CAV1 COL15A1	FOSL2 FOSB FLI1 HIF1A PBX1

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

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

The cell-type specific marker peaks and profiles derived from the reference samples were integrated to the EPIC deconvolution tool (17,25). We will refer to this ATAC-Seq deconvolution framework as EPIC-ATAC.

To test the accuracy of EPIC-ATAC predictions, we first collected PBMCs from five healthy donors. In each donor, half of the cells was used to generate a bulk ATAC-Seq dataset and the other half was used to determine the cellular composition of each sample, *i.e.,* the proportions of monocytes, B cells, CD4+ T cells, CD8+ T cells, NK cells and dendritic cells, by multiparametric flow cytometry (Figure 3A, see Materials and methods). We then applied EPIC-ATAC to the bulk ATAC-Seq data. The predicted cell fractions are consistent with the cell fractions obtained by flow cytometry (Figure 3B, Pearson correlation coefficient of 0.78 and root mean squared error (RMSE) of 0.10).

As a second validation, we applied EPIC-ATAC to pseudo-bulk PBMC samples (referred to as the PBMC pseudobulk dataset, generated using three publicly available PBMC scATAC-Seq datasets

190 (13,61,62), see Material and methods). A high correlation (0.91) between EPIC-ATAC predictions and
191 true cell-type proportions and a low RMSE (0.05) were observed for this dataset (Figure 3C).
192 The accuracy of the predictions obtained with EPIC-ATAC was then compared with the accuracy of
193 other deconvolution approaches which could be used with our reference profiles and marker peaks
194 (Figure 3D-E). To this end, we considered both the DeconPeaker method (20) originally developed for
195 bulk ATAC-Seq as well as several algorithms developed for bulk RNA-Seq (CIBERSORTx (19),
196 QuanTIseq (21), ABIS (18), and MCPcounter (22)). To enable meaningful comparison across the cell
197 types considered in this work and use the method initially developed for bulk RNA-Seq
198 deconvolution, the marker peaks and profiles derived in this work were used in each of these
199 methods. DeconPeaker and CIBERSORTx include the option to define cell-type specific markers and
200 profiles from a set of reference samples. We thus fed our ATAC-Seq samples collection to both
201 algorithms and used the resulting profiles and marker peaks to perform bulk ATAC-Seq
202 deconvolution. The resulting predictions are referred to as DeconPeaker-Custom and CIBERSORTx-
203 Custom.
204 Many tools displayed high correlation and low RMSE values, similar to those of EPIC-ATAC, and no
205 single tool consistently outperformed the others (Figure 3D-E, Supplementary Figure 2A-C). The fact
206 that our marker peaks and reference profiles could be used with EPIC-ATAC and other existing tools
207 demonstrates their broad applicability.
208 Predictions accuracies were also evaluated in each cell type separately. Since the number of samples
209 was low in each dataset, samples from both datasets were combined for this analysis. EPIC-ATAC
210 demonstrated good accuracies across cell types with RMSE values ranging from 0.02 for B cells to
211 0.13 for NK cells (Supplementary Figure 3). As expected, predictions with all tools were more
212 accurate for frequent cell types with well-characterized markers (*e.g.,* CD8/CD4 T cells, B cells)
213 compared to less frequent cell types (*e.g.,* NK cells, dendritic cells) (Supplementary Figure 2 and 3).
214 Note that MCPcounter is a marker-based method that derives cell-type specific scores which cannot

215 be compared between cell types. This method was thus only included in the benchmark considering
216 each cell type separately.
217

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

220 We evaluated the ability of the EPIC-ATAC framework to predict not only immune and stromal cells
221 proportions but also the proportion of cells for which reference profiles are not available (*i.e.,*
222 uncharacterized cells). For this purpose, we considered two previously published scATAC-Seq
223 datasets containing basal cell carcinoma and gynecological cancer samples (13,63). We generated
224 two pseudobulk datasets by averaging the chromatin accessibility signal across all cells of each
225 sample (see Material and methods). Applying EPIC-ATAC to both datasets shows that this framework
226 is able to simultaneously predict the proportions of both uncharacterized cells and immune, stromal
227 and vascular cells (Figure 4A). In these cancer samples, the proportion of uncharacterized cells can be
228 seen as a proxy of the proportion of cancer cells.
229 As for the PBMC datasets, we compared EPIC-ATAC performances to other existing deconvolution
230 tools. For both datasets, EPIC-ATAC led to the highest performances and was the only method to
231 accurately predict the proportion of uncharacterized cells (Figure 4B, Supplementary Figure 4 and 5).
232 Although quanTIseq also allows users to perform such predictions, the method resulted in lower
233 correlation and higher RMSE values when comparing the estimated and true proportions of the
234 uncharacterized cells (Figure 4B, Supplementary Figure 4).
235 In the EPIC-ATAC and quanTIseq frameworks, predictions correspond to absolute cell-type fraction,
236 *i.e.,* proportions of all cells present in the bulk, while the estimations obtained from the other tools
237 correspond to relative cell fractions, *i.e.,* proportions of cells present in the reference profiles
238 (CIBERSORTx, DeconPeaker) or to scores with arbitrary units (ABIS, MCPcounter). We thus conducted
239 a second benchmark excluding the predictions of uncharacterized cell fractions and rescaling both
240 estimations and true proportions to sum to 1 (see Material and methods). EPIC-ATAC outperformed

241 most of the other methods also when excluding the uncharacterized cells (Figure 4C, Supplementary
242 Figure 4 and 5).
243 Supplementary Figure 6 reports the performances of each tool when considering each cell type
244 separately. Overall, EPIC-ATAC showed comparable or higher correlation and lower RMSE values
245 when compared to the other deconvolution tools.
246

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

249 To explore the limitations of ATAC-Seq deconvolution, we next evaluated whether EPIC-ATAC could
250 predict the proportions of T-cell subtypes. To this end, we considered naive and non-naive CD8+ as
251 well as naïve, helper/memory and T regulatory CD4+ T cells. We redefined our list of cell-type specific
252 marker peaks and reference profiles including also these five T-cell subtypes (Supplementary Tables
253 8-9, Supplementary Figure 7A) and observed that the markers were conserved in external data
254 (Supplementary Figure 7B). The annotations of the markers associated to the T-cell subtypes are
255 available in Supplementary Tables 10-13.
256 We capitalized on the more detailed cell-type annotation of the PBMC datasets as well as the basal
257 cell carcinoma dataset to evaluate the EPIC-ATAC prediction of cell-subtype fractions using these
258 updated markers and profiles. Overall, the correlations observed between the predictions and true
259 proportions of T cells decreased when considering T-cell subtypes rather than CD4+ and CD8+ cell
260 types only (Figure 5A). In particular, low accuracies were obtained for helper/memory CD4+ and
261 naïve T-cell subtypes (Figure 5B). Similar results were obtained using other deconvolution tools
262 (Supplementary Figure 8).
263

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

266 We applied EPIC-ATAC to a breast cancer cohort of 42 breast ATAC-Seq samples including samples
267 from two breast cancer subtypes, *i.e.,* 35 oestrogen receptor (ER)-positive human epidermal growth
268 factor receptor 2 (HER2)-negative (ER+/HER2-) samples and 7 triple negative (TN) tumors (64). No cell
269 sorting was performed in parallel to the chromatin accessibility sequencing. We thus used EPIC-ATAC
270 to estimate cell-type proportions. We observed a higher proportion of T cells, B cells, NK cells and
271 macrophages in the TN samples in comparison to ER+/HER2- samples (Figure 6A). We then compared
272 the cellular composition of ER+/HER2- subgroups identified in the original study (clusters CA-A, CA-B
273 and CA-C). A higher infiltration of T and B cells was observed in cluster CA-C and higher proportions
274 of endothelial cells and fibroblasts were observed in cluster CA-B (Figure 6B). These predictions are
275 consistent with the infiltration level estimations reported in the original publication, although no
276 differences in macrophages infiltration was observed between the ER+/HER2- subgroups in our case
277 (64).
278

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

281 We finally compared the accuracy of EPIC when applied on ATAC-Seq data and on RNA-Seq data. For
282 this purpose, we used the 10X multiome PBMC dataset (62) which provides for each cell both its
283 chromatin accessibility profile and its gene expression profile and simulated 100 pseudobulks with
284 diverse cellular compositions (see Material and methods). We used EPIC-ATAC to perform ATAC-Seq
285 based deconvolution on the chromatin accessibility levels of the peaks and the original EPIC tool to
286 perform standard RNA-seq deconvolution on the gene expression levels. ATAC-Seq peaks can also be
287 aggregated, based on peak distances to each gene, into gene activity (GA) variables as proxy for gene
288 expression. We thus applied the GA transformation to the 10x multiome PBMC dataset and
289 performed GA-based RNA deconvolution using the original EPIC tool (See Material and methods).
290 Figure 7 shows that EPIC-ATAC performs similarly to the EPIC RNA-seq based deconvolution and
291 outperforms the GA-based RNA deconvolution. The lower performances of GA based RNA

292 deconvolution could be explained by the fact that GA features, by construction, do not perfectly
293 match the transcriptomic data.
294

295 ## Discussion

296 Bulk chromatin accessibility profiling of biological tissues like tumors represents a reliable and
297 affordable technology to map the activity of gene regulatory elements across multiple samples in
298 different conditions. Here, we collected ATAC-Seq data from pure cell populations covering major
299 immune and non-immune cancer-relevant cell types from diseased, stimulated and healthy samples.
300 This enabled us to identify reliable cell-type specific marker peaks and chromatin accessibility profiles
301 for both PBMC and solid tumor sample deconvolution. We integrated these data in the EPIC
302 deconvolution framework to accurately predict the fraction of both malignant and non-malignant cell
303 types from bulk tumor ATAC-Seq samples.
304 In cases where specific cell types are expected in a sample but are not part of our list of reference
305 profiles (*e.g.,* neuronal cells in brain tumors), custom marker peaks and reference profiles can be
306 provided to EPIC-ATAC to perform cell-type deconvolution and we provide the code to generate such
307 markers and profiles based on ATAC-Seq data from sorted cells, following the approach developed in
308 this work (Figure 1, see Code availability).
309 Solid tumors contain large and heterogeneous fractions of cancer cells for which it is
310 challenging to build reference profiles. To our knowledge this work provides the first benchmark of
311 deconvolution tools adapted to ATAC-Seq data in the context of solid tumor samples. We show that
312 the EPIC-ATAC framework, in contrast to other existing tools, allows users to accurately predict the
313 proportion of cells not included in the reference profiles (Figure 4 and Supplementary Figure 4).
314 These uncharacterized cells can include cancer cells but also other non-malignant cells. Since the
315 major cell types composing TMEs were included in our reference profiles, the proportion of
316 uncharacterized cells approximates the proportion of the cancer cells in most cases.

317 The pseudobulk approach provides unique opportunities to design benchmarks with known
318 cell-type proportions but also comes with some limitations. Indeed, pseudobulks are generated from
319 single-cell data which are noisy and whose cell-type annotation is challenging in particular for closely
320 related cell types. These limitations might lead to chromatin accessibility profiles that deviates from
321 true bulk data and errors in the true cell-type proportions. For this reason, we anticipate that the
322 newly generated benchmarking PBMCs dataset with ground truth cell proportions obtained by flow
323 cytometry will nicely complement pseudobulk from scATAC-Seq data in future benchmarks of ATAC-
324 Seq deconvolution. The qualitative evaluation of our method on true bulk ATAC-Seq samples from
325 breast cancer patients and the observation of similar immune compositions in TN and ER+/HER2-
326 samples as the ones identified in the original paper (Figure 6) further support the accuracy of EPIC-
327 ATAC to deconvolve bulk ATAC-Seq data, without requiring additional scATAC-Seq data which are not
328 always available for all cancer types.

329 Overall the evaluation of the EPIC-ATAC deconvolution resulted in an average absolute error
330 of 7% across cell types. This number is consistent with previous observations in RNA-Seq data
331 deconvolution (17). Considering this uncertainty, the quantification of low frequency populations
332 remains challenging (65). While the estimated proportions of these populations by EPIC-ATAC are
333 low (*e.g.,* dendritic cells), comparing such estimations across samples should be performed with care
334 due to the uncertainty of the predictions.

335 Another limitation of cell-type deconvolution is often reached when closely related cell types
336 are considered. In the reference-based methods used in this study, this limit was reached when
337 considering T-cell subtypes in the reference profiles (Figure 5 and Supplementary figure 8). We thus
338 recommend to use the EPIC-ATAC framework using the markers and reference profiles based on the
339 major cell populations. We additionally provide the marker peaks of the T-cell subtypes which could
340 be used to build cell-type specific chromatin accessibility signatures or perform "peak set enrichment
341 analysis" similarly to gene set enrichment analysis (GSEA, (66)). Such application could be useful for

342 the annotation of scATAC-Seq data, which often relies on matched RNA-Seq data and for which there
343 is a lack of markers at the peak level (67).

344 Another possible application of our marker peaks relies on their annotation (Figure 2G,
345 Supplementary Tables 4-5), which could be used to expand the list of genes and CBPs associated to
346 each cell type or subtype. For example, the neutrophils marker peaks were enriched for motifs of TFs
347 such as SPI1 (Supplementary Table 4), which was not listed in the neutrophil genes in the databases
348 used for annotation but has been reported in previous studies as involved in neutrophils
349 development (68). The annotations related to the set of major cell types and T-cell subtypes are
350 provided in Supplementary Tables 4-5 and 10-11. Finally, the annotation of marker peaks highlighted
351 pathways involved in immune responses to tumoral environments (Figure 2G). Examples of these
352 pathways are the toll-like receptor signaling pathway involved in pathogen-associated and
353 recognition of damage-associated molecular patterns in diverse cell types including B and T cells
354 (69,70), glucan metabolic processes which are known to be related to trained immunity which can
355 lead to anti-tumor phenotype in neutrophils (71) or the Fc-receptor signaling observed in NK cells
356 (72,73). These observations suggest that our marker peaks contain regulatory regions not only
357 specific to cell types but also adapted to the biological context of solid tumors.

358

359 Conclusion

360 In this work, we identified biologically relevant cell-type specific chromatin accessibility markers and
361 profiles for all major cancer-relevant cell types. We capitalized on these markers and profiles to
362 predict cell-type proportions from bulk PBMC and solid tumor ATAC-Seq data
363 (https://github.com/GfellerLab/EPIC-ATAC). Evaluated on diverse tissues, EPIC-ATAC shows reliable
364 predictions of immune, stromal, vascular and cancer cell proportions. With the expected increase of
365 ATAC-Seq studies in cancer, the EPIC-ATAC framework will enable researchers to deconvolve bulk
366 ATAC-Seq data from tumor samples to support the analysis of regulatory processes underlying tumor
367 development, and correlate the TME composition with clinical variables.

368

369 # Materials and methods

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

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

372 We collected pure ATAC-Seq samples from 12 studies. The data include samples from (i) ten major
373 immune, stromal and vascular cell types (B (35,39,40), CD4+ (35,40–43), CD8+ (35,39–41,43), natural
374 killer (NK) (35,39), dendritic (DCs) cells (39,41,44), macrophages (40,41), monocytes (35,39,40,44,45)
375 and neutrophils (46,47) as well as fibroblasts (41,48) and endothelial (41,49) cells (See Figure 2A),
376 and (ii) eight tissues from distinct organs (*i.e* bladder, breast, colon, liver, lung, ovary, pancreas and
377 thyroid) from the ENCODE data (50,51). The list of the samples and their associated metadata
378 (including cell types and accession number of the study of origin) is provided in Supplementary Table
379 1. To limit batch effects, the samples were reprocessed homogeneously from the raw data (fastq
380 files) processing to the peak calling. For that purpose, raw fastq files were collected from GEO using
381 the SRA toolkit and the PEPATAC framework (74) was used to process the raw fastq files based on
382 the following tool: trimmomatic for adapter trimming, bowtie2 (with the PEPATAC default
383 parameters) for reads pre-alignment on human repeats and human mitochondrial reference
384 genome, bowtie2 (with the default PEPATAC parameters: *--very-sensitive -X 2000*) for alignment on
385 the human genome (hg38), samtools (PEPATAC default parameters: *-q 10*) for duplicates removal and
386 MACS2 (75) (PEPATAC default parameters: *--shift -75 --extsize 150 --nomodel --call-summits --*
387 *nolambda --keep-dup all -p 0.01*) for peak calling in each sample. After alignment, reads mapping on
388 chromosome M were excluded. TSS enrichment scores were computed for each sample and used to
389 filter out samples with low quality (criteria of exclusion: TSS score < 5) (See Supplementary Table 1
390 containing the TSS score of each sample). 789 samples (including 564 from our ten reference cell-
391 types) had a TSS score > 5.

392

393 **Generation of a consensus set of peaks**

394 Peak calling was performed in each sample individually. Peaks were then iteratively collapsed to
395 generate a set of reproducible peaks. For each cell type, peaks collapse was performed adapting the
396 iterative overlap peak merging approach proposed in the PEPATAC framework. A first peaks collapse
397 was performed at the level of each study of origin, *i.e.,* if peaks identified in distinct samples
398 overlapped (minimum overlap of 1bp between peaks), only the peak with the highest peak calling
399 score was kept. Also, only peaks detected in at least half of the samples of each study were
400 considered for the next step. If a study had only two samples, only peaks detected in both samples
401 were considered. After this first selection, a second round of peaks collapse was performed at the
402 cell-type level to limit batch effects in downstream analyses. For each cell type, only peaks detected
403 in all the studies of origin were considered. The final list of peaks was then generated by merging
404 each set of reproducible peaks. Peaks located on chromosome Y were excluded from the rest of the
405 analyses. ATAC-Seq counts were retrieved for each sample and each peak using featureCounts (76).
406

407 **Identification of cell-type specific markers**

408 **Differential accessibility analysis**

409 To identify cell-type specific markers, we split the samples collection in ten folds (created with the
410 *create_folds* function from the R package splitTools (77)). For each fold, we performed pairwise
411 differential accessibility analysis across the ten cell types considered in the reference samples as well
412 as the ENCODE samples from diverse organs. The differential analysis was performed using limma
413 ((78), version 3.56.2). Effective library sizes were computed using the method of trimmed mean of
414 M-values (TMM) from the edgeR package in R ((79), version 3.42.4). Due to differences of library size
415 across all samples collected, we used voom from the limma package (80) to transform the data and
416 model the mean-variance relationship. Finally, a linear model was fitted to the data to assess the
417 differential accessibility of each peak across each pair of cell types. To identify our marker peaks, all
418 peaks with log2 fold change higher than 0.2 were selected and ranked by their maximum adjusted *p*-
419 value across all pairwise comparisons. The top 200 features (with the lowest maximum adjusted *p*-

420 value) were considered as cell-type specific marker peaks. The marker peaks identified in at least
421 three folds were considered in the final list of marker peaks.
422
423 **Marker peaks filtering**
424 Modules of open chromatin regions accessible in all (universal modules) or in specific human tissues
425 have been identified in the study Zhang *et al.* (2). These regions were used to refine the set of marker
426 peaks and exclude peaks with residual accessibility in other cell types than those considered for
427 deconvolution. More precisely, for immune, endothelial and fibroblasts specific peaks, we filtered
428 out the peaks overlapping the universal modules as well as the tissue specific modules except the
429 immune (modules 8 to 25), endothelial (modules 26 to 35) and stromal related modules (modules 41
430 to 49 and 139-150) respectively. As a second filtering step, we retained markers exhibiting the
431 highest correlation patterns in tumor bulk samples from different cancer types, *i.e.,* The Cancer
432 Genome Atlas (TCGA) samples (5). We used the Cancer Genomics Cloud (CGC) (81) to retrieve the
433 ATAC-Seq counts for each marker peaks in each TCGA sample (using *featureCounts*). For each set of
434 cell-type specific peaks, we identified the most correlated peaks using the *findCorrelation* function of
435 the caret R package ((82), version 6.0-94) with a correlation cutoff value corresponding to the 90th
436 percentile of pairwise Pearson correlation values.
437
438 **Evaluation of the study of origin batch effect**
439 To identify potential batch effect issues, we run principal component analysis (PCA) based on the
440 cell-type specific peaks after normalizing ATAC-Seq counts using full quantile normalization (FQ-FQ)
441 implemented in the EDASeq R package (83) to correct for depth and GC biases. These data were used
442 to visualize the data in two-dimensional space running Uniform Manifold Approximation (UMAP)
443 based on the PBMC and TME markers (Figure 2B). We also run PCA and used the ten first principal
444 components to evaluate distances between samples and compute silhouette coefficients based on
445 the cell type and study of origin classifications.

446

447 **Building the reference profiles**

448 It has been previously demonstrated in the context of RNA-Seq based deconvolution approaches
449 (16,17) that the transcripts per million (TPM) transformation is appropriate to estimate cell fractions
450 from bulk mixtures. We thus normalized the ATAC-Seq counts of the reference samples using a TPM-
451 like transformation, *i.e.,* dividing counts by peak length, correcting samples counts for depth and
452 rescaling counts so that the counts of each sample sum to 10^6. We then computed for each peak the
453 median of the TPM-like counts across all samples from each cell type to build the reference profiles
454 of the ten cell types considered in the EPIC-ATAC framework (Figure 2C). In the EPIC algorithm,
455 weights reflecting the variability of each feature of the reference profile can be considered in the
456 constrained least square optimization. We thus also computed the inter-quartile range of the TPM-
457 like counts for each feature in each cell type. Two ATAC-Seq reference profiles are available in the
458 EPIC-ATAC framework: (i) a reference profile containing profiles for B cells, CD4+ T cells, CD8+ T cells,
459 NK, monocytes, dendritic cells and neutrophils to deconvolve PBMC samples, and (ii) a reference
460 profile containing profiles for B cells, CD4+ T cells, CD8+ T cells, NK, dendritic cells, macrophages,
461 neutrophils, fibroblasts and endothelial cells to deconvolve tumor samples. The reference profiles
462 are available in the EPICATAC R package and the reference profiles restricted to our cell-type specific
463 marker peaks are available in the Supplementary Tables 2 and 3.

464

465 **Assessing the reproducibility of the marker peaks signal in independent samples**

466 We evaluated the chromatin accessibility level of the marker peaks in samples that were not included
467 in the peak calling step. Firstly, we considered samples from two independent studies (84,85)
468 providing pure ATAC-Seq data for five immune cell types (*i.e.,* B, CD4+ T cells, CD8+ T cells,
469 Monocytes, Macrophages) (Figure 2D). To consider the other cell types, samples that were excluded
470 from the reference dataset due to a low TSS enrichment score were also considered in this validation
471 dataset (Supplementary Table 1). Secondly, we collected the data from a single-cell atlas chromatin

472 accessibility from human tissues and considered the cell types included in our reference data (2)
473 (Figure 2E). We used the cell-type annotations provided in the original study (GEO accession number:
474 GSE184462). The Signac R package ((58), 1.9.0) was used to extract fragments counts for each cell
475 and each marker peak and the ATAC-Seq signal of each marker peak was averaged across all cells of
476 each cell type.
477
478 **Annotation of the marker peaks**
479 The cell-type specific markers were annotated using ChIPseeker R package ((53), version 1.34.1) and
480 the annotation from *TxDb.Hsapiens.UCSC.hg38.knownGene* in R to identify the regions in which the
481 marker peaks are (*i.e.,* promoter, intronic regions, etc.) and ChipEnrich to associate each peak to the
482 nearest gene TSS (54). The nearest genes identified were then compared to cell-type marker genes
483 listed in the PanglaoDB (55) and CellMarker databases (56). PanglaoDB provides an online interface
484 to explore a large collection on single-cell RNA-Seq data as well as a community-curated list of cell-
485 type marker genes. CellMarker is a database providing a large set of curated cell-type markers for
486 more than 400 cell types in human tissues retrieved from a large collection of single-cell studies and
487 flow cytometry, immunostaining or experimental studies. ChipEnrich was also used to perform gene
488 set enrichment and identify for each set of cell-type specific peaks potential biological pathways
489 regulated by the marker peaks. The enrichment analysis was performed using the *chipenrich* function
490 (*genesets = "GOBP" , locusdef = "nearest_tss"*) from the chipenrich R package (v2.22.0).
491 Chromatin accessibility peaks can also be annotated for chromatin binding proteins (CBPs) such as
492 transcription factors (TFs), whose potential binding in the peak region is reported in databases. In our
493 study we chose the JASPAR2022 (57) database and the ReMap database (60).
494 Using the JASPAR2022 database, we assessed, for each cell type, whether the cell-type specific
495 marker peaks were enriched in specific TFs motifs using two TFs enrichment analysis frameworks:
496 Signac (58) and MonaLisa (59). For the MonaLisa analysis, the cell-types specific markers peaks were
497 categorized in bins of sequences, one bin per cell type (use of the *calcBinnedMotifEnrR* function). To

498 test for an enrichment of motifs, the sequences of each bin were compared to a set of background
499 peaks with similar average size and GC composition obtained by randomly sampling regions in all the
500 peaks identified from the reference dataset. The enrichment test was based on a binomial test. For
501 the Signac analysis, we used the *FindMotif* function to identify over-represented TF motifs in each set
502 of cell-type specific marker peaks (query). This function used a hypergeometric test to compare the
503 number of query peaks containing the motif with the total number of peaks containing the motif in
504 the background regions (matched to have similar GC content, region length and dinucleotide
505 frequencies as the query regions), corresponding in our case to the peaks called in the reference
506 dataset.
507 The ReMap database associates chromatin binding proteins (CBPs), including TFs, transcriptional
508 coactivators and chromatin-remodeling factors, to their DNA binding regions based on DNA-binding
509 experiments such as chromatin immunoprecipitation followed by sequencing (ChIP-seq). For each
510 association of a CBP to its binding region, the cell type in which the binding has been observed is
511 reported in the ReMap database (biotype). We used the ReMapEnrich R package (version 0.99) to
512 test if the cell-type specific marker peaks are significantly enriched in CBPs-binding regions listed in
513 the Remap 2022 catalog. We considered the non-redundant peaks catalog from Remap 2022,
514 containing non-redundant binding regions for each CBP in each biotype. Similarly to the previously
515 mentioned enrichment methods, we chose the consensus peaks called in the reference samples as
516 universe for the enrichment test. Note that, for each cell type, an enrichment was retained only if the
517 biotype in which the CBP-regions were identified matched the correct cell-type.
518
519 **Running EPIC-ATAC on bulk ATAC-Seq data**
520 The samples used to generate the reference profiles were aligned using the hg38 reference genome.
521 To assure the compatibility of any input bulk ATAC-Seq dataset with the EPIC-ATAC marker peaks and
522 reference profiles, we provide an option to lift over hg19 datasets to hg38 (use of the liftOver R
523 package). Subsequently, the features of the input bulk matrix are matched to our reference profiles

524 features. To match both sets of features, we determine for each peak of the input bulk matrix the
525 distance to the nearest peak in the reference profiles peaks. Overlapping regions are retained and
526 the feature IDs are matched to their associated nearest peaks. If multiple features are matched to
527 the same reference peak, the counts are summed. In RNA-Seq based deconvolution, EPIC uses an
528 estimation of the amount of mRNA in each reference cell type to derive cell proportions. For the
529 ATAC-Seq based deconvolution these values were set to 1 to give similar weights to all cell-types
530 quantifications.
531

532 **Datasets used for the evaluation of ATAC-Seq deconvolution**

533 **PBMCs ATAC-Seq data from healthy donors**

534 ***Peripheral blood mononuclear cell (PBMC) isolation***

535 Venous blood from five healthy donors was collected at the local blood transfusion center of Geneva
536 in Switzerland, under the approval of the Geneva University Hospital's Institute Review Board, upon
537 written informed consent and in accordance with the Declaration of Helsinki. PBMCs were freshly
538 isolated by Lymphoprep (Promega) centrifugation (1800 rpm, 20 minutes, without break, room
539 temperature). Red blood cell lysis was performed using red blood lysis buffer (Qiagen) and platelets
540 were removed by centrifugation (1000 rpm, 10 minutes without break, room temperature). Cells
541 were counted and immediately used.
542

543 ***Flow cytometry***

544 Immune cell populations were identified using multiparameter flow cytometry and the following
545 antibodies: FITC anti-human CD45RA (HI100, Biolegend), PerCP-Cyanine5.5 anti-human CD19 (H1B19,
546 Biolegend), PE anti-human CD3 (SK7, Biolegend), PE-Dazzle anti-human CD14 (M0P9, BD
547 Biosciences), PE-Cyanine7 anti-human CD56 (HCD56, Biolegend), APC anti-human CD4 (RPA-T4,
548 Biolgend), APC-Cyanine7 anti-human CCR7 (G043H7, Biolegend), Brilliant Violet 421 anti-human CD8
549 (RPA-T8, Biolegend), Brilliant Violet 510 anti-human CD25 (BC96, Biolegend), Brilliant Violet 711 anti-

550 human CD16 (3G8, Biolegend), Brilliant Violet 786 anti-human CD127 (A019D5, Biolegend), Ultra-
551 Brilliant Violet anti-human CD45 (HI30, BD Biosciences), FITC anti-human Celc9a (8F9, Miltenyi) , PE
552 anti-human XCR1 (S15046E, Biolegend), PE-Dazzle anti-human BDCA-2 (201A, Biolegend), APC anti-
553 human BDCA-3 (AD5-14H12, Miltenyi), Brilliant Violet 421 anti-human CD3 (UCHT1, Biolegend),
554 Brilliant Violet 421 anti-human CD14 (M5E2, BD Pharmingen), Brilliant Violet 421 anti-human CD19
555 (SJ25C1, Biolegend), Brilliant Violet 510 anti-human BDCA-1 (L161, Biolegend), Brilliant Violet 650
556 anti-human CD11c (3.9, Biolegend), Brilliant Violet 711 anti-human CD11c (N418, Biolegend) and
557 Brilliant Violet 711 anti-human HLA-DR (L243, Biolegend). Dead cells were excluded using the Zombie
558 UV™ Fixable Viability Kit (Biolegend). Intracellular staining was performed after fixation and
559 permeabilization of the cells with the FoxP3 Transcription Factor Staining Buffer Set (00-5523-00,
560 Invitrogen) using Alexa 700 anti-human FoxP3 antibody (259D/C7, BD Biosciences). Data were
561 acquired on LSRFortessa flow cytometer and analysed using FlowJo software (v10.7.1).
562
563 ***Cell preparation for ATAC-Sequencing***
564 50000 CD45+ cells were sorted from total PBMCs using anti-human Ultra-Brilliant Violet (BUV395)
565 CD45 (HI30, BD Biosciences) with a FACSAria II (Becton Dickinson) and were collected in PBS with
566 10% Foetal Bovine Serum (FBS). Cell pellets were resuspended in cold lysis buffer (10mM Tris-Cl pH
567 7.4, 10mM NaCl, 3mM MgCl2, 0,1% NP40 and water) and immediately centrifuged at 600g for 30min
568 at 4°C. Transposition reaction was performed using the Illumina Tagment DNA Enzyme and Buffer kit
569 (20034210, Illumina) and transposed DNA was eluted using the MinElute PCR Purification Kit
570 (Qiagen). Libraries were generated by PCR amplification using indexing primers and NEBNext High-
571 Fidelity Master Mix (New England BioLabs) and were purified using AMPure XP beads (A63880,
572 Beckman Coulter). Libraries were quantified by a fluorometric method (QubIT, Life Technologies) and
573 their quality assessed on a Fragment Analyzer (Agilent Technologies). Sequencing was performed as a
574 paired end 50 cycles run on an Illumina NovaSeq 6000 (v1.5 reagents) at the Genomic Technologies

575 Facility (GTF) in Lausanne, Switzerland. Raw sequencing data were demultiplexed using the
576 bcl2fastq2 Conversion Software (version 2.20, Illumina).
577
578 ***Data processing***
579 The same steps as for the processing of the reference ATAC-Seq samples were followed. (See Pre-
580 processing of the ATAC-Seq datasets).
581
582 **ATAC-Seq pseudobulk data from PBMCs and cancer samples**
583 To evaluate the accuracy of our ATAC-Seq deconvolution framework, we generated pseudo-bulk
584 datasets from 5 single-cell datasets:
585 - **PBMC pseudobulk dataset:** combination of three single-cell datasets for PBMCs.
586 - Dataset 1 corresponds to a scATAC-Seq dataset obtained from Satpathy et al. (13)
587 (GEO accession number: GSE129785). This dataset contains FACS-sorted populations
588 of PBMCs. Since the cells of some cell types came from a unique donor, all the cells
589 of this dataset were aggregated to form one pseudobulk. Ground truth cell fractions
590 were obtained by dividing the number of cells in each cell type by the total number
591 of cells.
592 - Dataset 2 (included in the PBMC pseudobulk dataset) was retrieved from Granja *et*
593 *al.* (61) (GEO accession number GSE139369). B cells, monocytes, dendritic, CD8+,
594 CD4+ T, NK cells, neutrophils from healthy donors were considered. The neutrophil
595 cells came from a single donor. As for dataset 1, we thus aggregated all the cells to
596 generate one pseudobulk. Ground truth cell fractions were obtained by dividing the
597 number of cells in each cell type by the total number of cells.
598 - Dataset 3 (included in the PBMC pseudobulk dataset) corresponds to the 10X
599 multiome dataset of PBMC cells (62). Since these data come from one donor, one
600 pseudobulk sample was generated for this dataset. The pseudobulk was generated

601 by averaging the ATAC-Seq signal from all cells from the following cell types: B cells,
602 CD4+ T cells , CD8+ T cells, NK cells, Dendritic cells and monocytes.

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

609 - **Gynecological cancer dataset:** obtained from the study of Regner *et al.* (63) (GEO accession
610 number GSE173682). In this study, the authors performed scATAC-Seq on 11 gynecological
611 cancer samples from two tumor sites (*i.e* endometrium and ovary) and composed of immune
612 (B cells, NK and T cells grouped under the same cell-type annotation, macrophages, mast
613 cells), stromal (fibroblast, endothelial, smooth muscle) and cancer cells. Mast cells, smooth
614 muscle and cancer cells were considered as uncharacterized cells. Cell annotations were
615 retrieved from the original study.

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

619 For each dataset, raw fragments files were downloaded from the respective GEO accession numbers
620 and data were preprocessed using ArchR ((86), ArchR R package 1.0.2). Cells with TSS score below
621 four were removed. Doublets removal was performed using the *doubletsRemoval* function from
622 ArchR. To match as much as possible real bulk ATAC-seq data processing, peak calling was not
623 performed on each cell type or cell cluster as usually done in scATAC-Seq studies but using all cells
624 for each dataset from the PBMC pseudobulk data or grouping cells by sample for the Basal cell
625 carcinoma and Gynecological cancer datasets. Peak calling was performed using MACS2 within the
626 ArchR framework. Fragments counts were extracted using ArchR for each peak called to generate

627 single-cell peak counts matrices. These matrices were normalized using a TPM-like transformation,
628 *i.e.,* dividing counts by peak length and correcting samples counts for depth. Finally, for each peak,
629 the average of the normalized counts was computed across all the cells for each dataset from the
630 PBMC pseudobulk data and across all the cells of each sample for the Basal cell carcinoma and
631 Gynecological cancer datasets. Averaged data were then rescaled so that the sum of counts of each
632 sample sum to 10^6.
633
634 **Bulk ATAC-Seq data from a breast cancer cohort**
635 Bulk ATAC-Seq samples from a breast cancer cohort was obtained from Kumegawa *et al.* (64). These
636 data include 42 breast cancer samples which can be classified based on two features: (i) the breast
637 cancer subtype ER+/HER2- or triple negative, and (ii) the molecular classification provided by the
638 original study (CA-A, CA-B and CA-C). The ATAC-Seq raw counts and the samples metadata were
639 retrieved from figshare (87). As for the previously mentioned datasets, raw counts were normalized
640 using the TPM-like transformation prior to bulk deconvolution.
641
642 **<u>Benchmarking of the EPIC-ATAC framework against other existing deconvolution tools</u>**
643 The performances of the EPIC-ATAC framework were benchmarked against the following
644 deconvolution tools:
645 - quanTIseq (21) is a deconvolution tool using constrained least square regression to
646 deconvolve RNA-Seq bulk samples. No reference profiles are available in this framework to
647 perform ATAC-Seq deconvolution and quanTIseq does not provide the option to
648 automatically build reference profiles from pure bulk samples. quanTIseq was thus run using
649 the reference profiles derived in this work for the EPIC-ATAC framework and the *quanTIseq*
650 function from the quantiseqr R package (parameters: scaling set to 1 for all cell types and
651 method set to "lsei").

652 • DeconPeaker (20) relies on SIMPLS, a variant of partial least square regression to perform
653 bulk RNA-Seq and bulk ATAC-Seq deconvolution. ATAC-Seq reference profiles are available in
654 this deconvolution framework however not all cell types considered in the EPIC-ATAC
655 framework are included in the DeconPeaker reference profiles. This tool was thus run using
656 different reference profiles: (i) the reference profiles derived in this work for the EPIC-ATAC
657 framework (corresponds to "DeconPeaker" or "DeconPeaker_ourmarkers" in our analyses),
658 and (ii) reference profiles automatically generated by DeconPeaker from the sorted
659 reference samples collected in this work (corresponds to "DeconPeaker_cust." in our
660 analyses). The results of DeconPeaker obtained using its original markers and profiles are
661 also provided for the cell types in common with the cell types considered in this work in
662 Supplementary Figures 3 and 5. Deconvolution was run using the deconvolution module
663 deconPeaker (using findctsps with the following parameter: --lib-strategy=ATAC-Seq).
664 DeconPeaker outputs cell-type proportions relative to the total amount of cells from the
665 reference cell types.
666 • CIBERSORTx (19) is a deconvolution algorithm based on linear support vector regression.
667 CIBERSORTx does not provide ATAC-Seq reference profiles, however it is possible to
668 automatically generate new profiles from a set of pure bulk samples. This tool was thus run
669 using different reference profiles: i) the reference profiles derived in this work for the EPIC-
670 ATAC framework (corresponds to "CIBERSORTx" or "CIBERSORTx_ourMarkers" in our
671 analyses), and ii) reference profiles automatically generated by CIBERSORTx from the sorted
672 reference samples collected in this work (corresponds to "CIBERSORTx_cust." in our
673 analyses). To run CIBERSORTx, we used the docker container provided by the authors of
674 CIBERSORTx on their website. The algorithm was run using the default options (i.e --absolute
675 FALSE, --rmbatchBmode FALSE and –rmbatchSmode FALSE), which results in cell-type
676 proportions relative to the total amount of cells from the reference cell types.

677 • ABIS (18) uses robust linear modeling to estimate cell-type proportions in bulk RNA-Seq
678 samples. No ATAC-Seq reference profiles are available in the deconvolution framework. ABIS
679 was run using the EPIC-ATAC reference profiles by using the *rlm* function from the MASS R
680 package (as performed in the deconvolute_abis function from the immunedeconv R package
681 (16) was used to quantify each cell type from the reference profiles. The cell-types
682 quantifications returned by this approach are in arbitrary units. To compare the estimations
683 and the true cell proportions, we scaled the estimations of each sample between 0 and 1 to
684 obtained relative proportions.
685 • MCPcounter (22): MCPcounter returns scores instead of cell type proportions. The scores
686 were obtained using the *appendSignatures* function from the MCPcounter R package by
687 providing the list of marker peaks specific to each cell type. The cell-type scores are not
688 comparable between cell type, MCPcounter was thus included only in the evaluation of the
689 performances in each cell type separately.
690
691 For all the tools, TPM-like data were used as input bulk samples for the deconvolution.
692 Since CIBERSORTx, ABIS and DeconPeaker do not predict proportions of uncharacterized cells, we
693 performed two benchmarking analyses: (i) including all cell types and (ii) excluding the cell types that
694 are absent from the reference profiles (uncharacterized cells) and rescaling the estimated and true
695 proportions of the immune cells, endothelial cells and fibroblasts so that their sum equals 1.
696
697 **Comparing deconvolution based on RNA-Seq, gene activity or peaks features.**
698 100 pseudobulks were generated from the 10X PBMC multiome dataset (62) based on 3000 cells for
699 each pseudobulk. Cell fractions were defined using the *rdirichlet* function from the *gtools* R package.
700 Three sets of features were extracted from the data, *i.e.,* gene expression features extracted from
701 the RNA-Seq layer, ATAC-Seq peaks and gene activity derived from the ATAC-Seq layer. The same
702 cells sampling was considered for each modality.

703 Gene activity features were extracted from the single-cell data using ArchR (1.0.2), which considers
704 distal elements and adjusts for large differences in gene size in the gene activity score calculation.
705 Gene activity pseudobulks were built by averaging the gene activity scores across all cells belonging
706 to the pseudobulk. For ATAC-Seq pseudobulk, peaks called using ArchR on all cells form the 10X
707 dataset were considered (see the method section "ATAC-Seq pseudobulk data from PBMCs and
708 cancer samples") and counts were averaged across all cells of each pseudobulk. For RNA-Seq
709 pseudobulks, counts were also averaged across all cells of each pseudobulk. All aggregated data were
710 depth normalized across each features to 10^6. Cell-type deconvolution was performed on each
711 pseudobulk using EPIC-ATAC on the peak matrix using our ATAC-Seq marker peaks and reference
712 profiles. The RNA-Seq and gene activity pseudobulks were deconvolved with EPIC.
713
714 **Code availability**
715 The code to download and preprocess publicly available ATAC-Seq samples as well as the code used
716 to identify our cell-type specific marker peaks and generate the reference profiles is available on
717 GitHub (https://github.com/GfellerLab/EPIC-ATAC_Manuscript). A README file is provided on the
718 GitHub repository with more details on how to use the code.
719 The code to perform ATAC-Seq deconvolution using the EPIC-ATAC framework is available as an R
720 package called EPICATAC and is available on GitHub (https://github.com/GfellerLab/EPIC-ATAC).
721
722 **Data availability**
723 The newly generated ATAC-Seq data have been deposited on Zenodo (doi:
724 10.5281/zenodo.8358719). The other data related to this work are available in the supplementary
725 tables and on the Zenodo deposit (doi: 10.5281/zenodo.8358719).
726
727 **Competing interests:**
728 The authors declare that they have no competing interests.

Acknowledgement:

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

Authors contributions:

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

Figure legends:

Figure 1: Graphical description of the identification of cell-type specific marker peaks and reference ATAC-Seq profiles included in the EPIC-ATAC framework. *1) 564 pure ATAC-Seq data of sorted cells were collected to build reference profiles for cancer-relevant cell populations. 2) Cell-type specific marker peaks were identified using differential accessibility analysis. 3) Markers with previously observed chromatin accessibility in human healthy tissues were then excluded. 4) For tumor bulk deconvolution, the set of remaining marker peaks was refined by selecting markers with correlated behavior in tumor bulk samples. 5) The cell-type specific marker peaks and reference profiles were finally integrated in the EPIC-ATAC framework to perform bulk ATAC-Seq deconvolution. Parts of this figure were created with BioRender.com.*

Figure 2: ATAC-Seq data from sorted cell populations reveal cell-type specific marker peaks and reference profiles. A) *Number of samples collected for each cell type. The colors correspond to the different studies of origin.* ***B)*** *Representation of the collected samples in 2D using UMAP based on the*

755 *PBMC markers (left) and TME markers (right). Colors correspond to cell types.* ***C)*** *Scaled averaged*
756 *chromatin accessibility of the cell-type specific marker peaks (rows) in each cell type (columns) in the*
757 *ATAC-Seq reference samples used to identify the marker peaks.* ***D)*** *Scaled averaged chromatin*
758 *accessibility of the marker peaks in external ATAC-Seq data from samples of pure cell types excluded*
759 *from the reference samples (see Material and Methods).* ***E)*** *Scaled averaged chromatin accessibility of*
760 *the marker peaks in an external scATAC-Seq dataset (Human Atlas* (2)*).* ***F)*** *Distribution of the marker*
761 *peak distances to the nearest transcription start site (TSS) (left panel) and the ChiPSeeker annotations*
762 *(right panel).* ***G)*** *Significance (-log10(q.value)) of pathways (columns) enrichment test obtained using*
763 *ChIP-Enrich on each set of cell-type specific marker peaks (rows). A subset of relevant enriched*
764 *pathways is represented. Colors of the names of the pathways correspond to cell types where the*
765 *pathways were found to be enriched. When pathways were significantly enriched in more than one*
766 *set of peaks, pathways names are written in bold.*
767
768 ***Figure 3: EPIC-ATAC accurately estimates immune cell fractions in PBMC ATAC-Seq samples. A)***
769 *Schematic description of the experiment designed to validate the ATAC-Seq deconvolution on PBMC*
770 *samples.* ***B)*** *Comparison between cell-type proportions predicted by EPIC-ATAC and the true*
771 *proportions in the PBMC bulk dataset. Symbols correspond to donors.* ***C)*** *Comparison between the*
772 *proportions of cell-types predicted by EPIC-ATAC and the true proportions in the PBMC pseudobulk*
773 *dataset. Symbols correspond to pseudobulks.* ***D)*** *Pearson correlation (left) and RMSE (right) values*
774 *obtained by each deconvolution tool on the PBMC bulk dataset. The EPIC-ATAC results are highlighted*
775 *in red.* ***E)*** *Pearson correlation (left) and RMSE (right) values obtained by each deconvolution tool on*
776 *the PBMC pseudobulk dataset. Parts of this figure (panel 1) were created with BioRender.com.*
777
778 ***Figure 4: EPIC-ATAC accurately predicts fractions of cancer and non-malignant cells in tumor***
779 ***samples. A)*** *Comparison between cell-type proportions estimated by EPIC-ATAC and true proportions*
780 *for the basal cell carcinoma (top) and gynecological (bottom) pseudobulk datasets. Symbols*

781 *correspond to pseudobulks.* ***B)*** *Pearson's correlation and RMSE values obtained for the deconvolution*
782 *tools included in the benchmark. EPIC-ATAC is highlighted in red.* ***C)*** *Same analyses as in panels B,*
783 *with the uncharacterized cell population excluded for the evaluation of the predictions accuracy. The*
784 *predicted and true proportions of the immune, stromal and vascular cell types were rescaled to sum*
785 *to 1.*
786
787 ***Figure 5: T cell subtypes quantification reveals the ATAC-Seq deconvolution limits for closely***
788 ***related cell types. A)*** *Comparison of the proportions estimated by EPIC-ATAC and the true proportions*
789 *for PBMC samples (PBMC experiment and PBMC pseudobulk samples combined) (top) and the basal*
790 *cell carcinoma pseudobulks (bottom). Predictions of the proportions of CD4+ and CD8+ T-cells were*
791 *obtained using the reference profiles based on the major cell types and subtype predictions using the*
792 *reference profiles including the T-cell subtypes.* ***B)*** *Pearson's correlation values obtained by EPIC-ATAC*
793 *in each cell type.*
794
795 ***Figure 6: EPIC-ATAC accurately infers the immune contexture in a bulk ATAC-Seq breast cancer***
796 ***cohort. A)*** *Proportions of different cell types predicted by EPIC-ATAC in the samples stratified based*
797 *on two breast cancer subtypes.* ***B)*** *Proportions of different cell types predicted by EPIC-ATAC in the*
798 *samples stratified based on three ER+/HER2- subgroups. Wilcoxon test p-values are represented at*
799 *the top of the boxplots.*
800
801 ***Figure 7: EPIC-ATAC performs similarly to EPIC RNA-seq based deconvolution and better than gene***
802 ***activity based deconvolution.*** *Pearson's correlation (left) and RMSE (right) values comparing the*
803 *proportions predicted by the ATAC-Seq deconvolution, the RNA-Seq deconvolution and the GA-based*
804 *RNA deconvolution and true cell-type proportions in the 100 pseudobulks simulated form the 10x*
805 *multiome PBMC dataset* (62). *Dots correspond to outlier pseudobulks.*
806

807 **Supplementary Figures:** Additional file named Supplementary_figures.pdf

808 **Supplementary Tables:**

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

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

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

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

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

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

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

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

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

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

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

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

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

826

827 **References:**

828

829 1. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome.
830 Nat Rev Genet. 2019 Jan 23;20(4):207–20.

831 2. Zhang K, Hocker JD, Miller M, Hou X, Chiou J, Poirion OB, et al. A single-cell atlas of chromatin
832 accessibility in the human genome. Cell. 2021;184(24):5985-6001.e19.

833 3. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin
834 for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and
835 nucleosome position. Nat Methods. 2013 Oct 6;10(12):1213–8.
836 4. Luo L, Gribskov M, Wang S. Bibliometric review of ATAC-Seq and its application in gene
837 expression. Brief Bioinform. 2022 Mar 7;
838 5. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin
839 accessibility landscape of primary human cancers. Science. 2018 Oct 26;362(6413).
840 6. Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, et al. An
841 improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues.
842 Nat Methods. 2017 Oct 1;14(10):959–62.
843 7. Grandi FC, Modi H, Kampman L, Corces MR. Chromatin accessibility profiling by ATAC-seq. Nat
844 Protoc. 2022 Apr 27;1518–52.
845 8. Fridman WH, Pagès F, Sautès-Fridman C, Galon J. The immune contexture in human tumours:
846 impact on clinical outcome. Nat Rev Cancer. 2012 Mar 15;12(4):298–306.
847 9. Fridman WH, Zitvogel L, Sautès-Fridman C, Kroemer G. The immune contexture in cancer
848 prognosis and treatment. Vol. 14, Nature Reviews Clinical Oncology. Nature Publishing Group;
849 2017. p. 717–34.
850 10. de Visser KE, Joyce JA. The evolving tumor microenvironment: From cancer initiation to
851 metastatic outgrowth. Cancer Cell. 2023 Mar 13;41(3):374–403.
852 11. Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, et al. Multiplex
853 single-cell profiling of chromatin accessibility by combinatorial cellular indexing. Science. 2015
854 May 22;348(6237):910–4.
855 12. Lareau CA, Duarte FM, Chew JG, Kartha VK, Burkett ZD, Kohlway AS, et al. Droplet-based
856 combinatorial indexing for massive-scale single-cell chromatin accessibility. Nat Biotechnol.
857 2019 Jun 24;37(8):916–24.
858 13. Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi F, McDermott GP, et al. Massively parallel

859 single-cell chromatin landscapes of human immune cell development and intratumoral T cell
860 exhaustion. Nat Biotechnol. 2019 Aug 1;37(8):925–36.
861 14. Avila Cobos F, Vandesompele J, Mestdagh P, De Preter K. Computational deconvolution of
862 transcriptomics data from mixed cell populations. Bioinformatics. 2018 Jun;34(11):1969–79.
863 15. Avila Cobos F, Alquicira-Hernandez J, Powell JE, Mestdagh P, De Preter K. Benchmarking of cell
864 type deconvolution pipelines for transcriptomics data. Nat Commun. 2020;11(1):1–14.
865 16. Sturm G, Finotello F, Petitprez F, Zhang JD, Baumbach J, Fridman WH, et al. Comprehensive
866 evaluation of transcriptome-based cell-type quantification methods for immuno-oncology.
867 Bioinformatics. 2019 Jul 5;35(14):i436–45.
868 17. Racle J, de Jonge K, Baumgaertner P, Speiser DE, Gfeller D. Simultaneous enumeration of
869 cancer and immune cell types from bulk tumor gene expression data. Elife. 2017 Nov 13;6.
870 18. Monaco G, Lee B, Xu W, Mustafah S, Hwang YY, Carré C, et al. RNA-Seq Signatures Normalized
871 by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. Cell Rep.
872 2019 Feb 5;26(6):1627-1640.e7.
873 19. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell
874 type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol.
875 2019;37(7):773–82.
876 20. Li H, Sharma A, Luo K, Qin ZS, Sun X, Liu H. DeconPeaker, a Deconvolution Model to Identify
877 Cell Types Based on Chromatin Accessibility in ATAC-Seq Data of Mixture Samples. Front
878 Genet. 2020 Jun 8;11.
879 21. Finotello F, Mayer C, Plattner C, Laschober G, Rieder Di, Hackl H, et al. Molecular and
880 pharmacological modulators of the tumor immune contexture revealed by deconvolution of
881 RNA-seq data. Genome Med. 2019 May 24;11:34.
882 22. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the
883 population abundance of tissue-infiltrating immune and stromal cell populations using gene
884 expression. Genome Biol. 2016 Oct 20;17:218.

885 23. Clarke J, Seol P, Clarke B. Statistical expression deconvolution from mixed tissue samples.
886 Bioinformatics. 2010;26(8):1043–9.
887 24. Gosink MM, Petrie HT, Tsinoremas NF. Electronically subtracting expression patterns from a
888 mixed cell population. Bioinformatics. 2007 Dec 15;23(24):3328–34.
889 25. Racle J, Gfeller D. EPIC: A tool to estimate the proportions of different cell types from bulk
890 gene expression data. In: Methods in Molecular Biology. Humana Press Inc.; 2020. p. 233–48.
891 26. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell
892 subsets from tissue expression profiles. Nat Methods. 2015 Mar 20;12(5):453–7.
893 27. Li T, Fu J, Zeng Z, Cohen D, Li J, Chen Q, et al. TIMER2.0 for analysis of tumor-infiltrating
894 immune cells. Nucleic Acids Res. 2020 May 22;48:W509–14.
895 28. Jimenez-Sanchez A, Cast O, Miller ML. Comprehensive benchmarking and integration of tumor
896 microenvironment cell estimation methods. Cancer Res. 2019 Dec 15;79(24):6238–46.
897 29. Gong T, Szustakowski JD. DeconRNASeq: a statistical framework for deconvolution of
898 heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics. 2013 Feb
899 21;29(8):1083–5.
900 30. Chakravarthy A, Furness A, Joshi K, Ghorani E, Ford K, Ward MJ, et al. Pan-cancer
901 deconvolution of tumour composition using DNA methylation. Nat Commun. 2018 Aug 13;9.
902 31. Teschendorff AE, Zhu T, Breeze CE, Beck S. EPISCORE: Cell type deconvolution of bulk tissue
903 DNA methylomes from single-cell RNA-Seq data. Genome Biol. 2020 Sep 4;21.
904 32. Arneson D, Yang X, Wang K. MethylResolver—a method for deconvoluting bulk DNA
905 methylation profiles into known and unknown cell contents. Commun Biol. 2020;3.
906 33. Zhang H, Cai R, Dai J, Sun W. EMeth: An EM algorithm for cell type decomposition based on
907 DNA methylation data. Sci Rep. 2021 Mar 11;11:5717.
908 34. Feng S, Calinawan A, Pugliese P, Wang P, Ceccarelli M, Petralia F, et al. Decomprolute : A
909 benchmarking platform designed for multiomics-based tumor deconvolution. bioRxiv. 2023;
910 35. Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, et al. Lineage-specific

911 and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution.
912 Nat Genet. 2016 Aug 15;48:1193–203.
913 36. Zeng W, Chen X, Duren Z, Wang Y, Jiang R, Wong WH. DC3 is a method for deconvolution and
914 coupled clustering from bulk and single-cell genomics data. Nat Commun. 2019 Oct
915 10;10:4613.
916 37. Burdziak C, Azizi E, Prabhakaran S, Pe'er D. A Nonparametric Multi-view Model for Estimating
917 Cell Type-Specific Gene Regulatory Networks. ArXiv. 2019;
918 38. Peng XL, Moffitt RA, Torphy RJ, Volmar KE, Yeh JJ. De novo compartment deconvolution and
919 weight estimation of tumor samples using DECODER. Nat Commun. 2019 Oct 18;10(1):4729.
920 39. Calderon D, Nguyen MLT, Mezger A, Kathiria A, Müller F, Nguyen V, et al. Landscape of
921 stimulation-responsive chromatin across diverse human immune cells. Nat Genet. 2019 Sep
922 30;51(10):1494–505.
923 40. Zhang P, Amarasinghe HE, Whalley JP, Tay C, Fang H, Migliorini G, et al. Epigenomic analysis
924 reveals a dynamic and context-specific macrophage enhancer landscape associated with
925 innate immune activation and tolerance. Genome Biol. 2022 Jun 24;23(1):136.
926 41. Liu Q, Zaba L, Satpathy AT, Longmire M, Zhang W, Li K, et al. Chromatin accessibility
927 landscapes of skin cells in systemic sclerosis nominate dendritic cells in disease pathogenesis.
928 Nat Commun. 2020 Nov 17;11(1):5843.
929 42. Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, et al. Enhancer connectome
930 in primary human cells identifies target genes of disease-associated DNA elements. Nat
931 Genet. 2017 Sep 15;49(11):1602–12.
932 43. Giles JR, Manne S, Freilich E, Oldridge DA, Baxter AE, George S, et al. Human epigenetic and
933 transcriptional T cell differentiation atlas for identifying functional T cell-specific enhancers.
934 Immunity. 2022 Mar;55(3):557-574.e7.
935 44. Leylek R, Alcántara-Hernández M, Granja JM, Chavez M, Perez K, Diaz OR, et al. Chromatin
936 Landscape Underpinning Human Dendritic Cell Heterogeneity. Cell Rep. 2020 Sep

937 22;32(12):108180.
938 45. Trizzino M, Zucco A, Deliard S, Wang F, Barbieri E, Veglia F, et al. EGR1 is a gatekeeper of
939 inflammatory enhancers in human macrophages. Sci Adv. 2021 Jan 13;7(3).
940 46. Ram-Mohan N, Thair SA, Litzenburger UM, Cogill S, Andini N, Yang X, et al. Profiling chromatin
941 accessibility responses in human neutrophils with sensitive pathogen detection. Life Sci
942 Alliance. 2021 Jun 18;4(8).
943 47. Perez C, Botta C, Zabaleta A, Puig N, Cedena M-T, Goicoechea I, et al. Immunogenomic
944 identification and characterization of granulocytic myeloid-derived suppressor cells in
945 multiple myeloma. Blood. 2020 Jul 9;136(2):199–209.
946 48. Ge X, Frank-Bertoncelj M, Klein K, McGovern A, Kuret T, Houtman M, et al. Functional
947 genomics atlas of synovial fibroblasts defining rheumatoid arthritis heritability. Genome Biol.
948 2021 Aug 25;22(1):247.
949 49. Xin J, Zhang H, He Y, Duren Z, Bai C, Chen L, et al. Chromatin accessibility landscape and
950 regulatory network of high-altitude hypoxia adaptation. Nat Commun. 2020 Oct 1;11(1):4928.
951 50. The ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, et al.
952 Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020
953 Jul 29;583:699–710.
954 51. Rozowsky J, Gao J, Borsari B, Guigó R, Gingeras TR, Gerstein M, et al. The EN-TEx resource of
955 multi-tissue personal epigenomes & variant-impact models. Cell. 2023;186(7):1493-1511.e40.
956 52. Qiu Y, Wang J, Lei J, Roeder K. Identification of cell-type-specific marker genes from co-
957 expression patterns in tissue samples. Bioinformatics. 2021 Oct 11;37(19):3228–34.
958 53. Yu G, Wang LG, He QY. ChIPseeker: an R/Bioconductor package for ChIP peak annotation,
959 comparison and visualization. Bioinformatics. 2015 Jul 15;31(14):2382–3.
960 54. Welch RP, Lee C, Imbriano PM, Patil S, Weymouth TE, Smith RA, et al. ChIP-Enrich: gene set
961 enrichment testing for ChIP-seq data. Nucleic Acids Res. 2014 Sep 1;42(13):e105.
962 55. Franzén O, Gan LM, Björkegren JLM. PanglaoDB: a web server for exploration of mouse and

963 human single-cell RNA sequencing data. Database. 2019;2019:baz046.
964 56. Hu C, Li T, Xu Y, Zhang X, Li F, Bai J, et al. CellMarker 2.0: an updated database of manually
965 curated cell markers in human/mouse and web tools based on scRNA-seq data. Nucleic Acids
966 Res. 2023 Jan 6;51(D1):D870–6.
967 57. Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Berhanu Lemma R, Turchi L, Blanc-
968 Mathieu R, et al. JASPAR 2022: the 9th release of the open-access database of transcription
969 factor binding profiles. Nucleic Acids Res. 2022 Jan 7;50(D1):D165–73.
970 58. Stuart T, Srivastava A, Madad S, Lareau CA, Satija R. Single-cell chromatin state analysis with
971 Signac. Nat Methods. 2021 Nov 1;18:1333–41.
972 59. Machlab D, Burger L, Soneson C, Rijli FM, Schübeler D, Stadler MB. monaLisa: an
973 R/Bioconductor package for identifying regulatory motifs. Bioinformatics. 2022 Feb
974 22;38(9):2624–5.
975 60. Hammal F, De Langen P, Bergon A, Lopez F, Ballester B. ReMap 2022: a database of Human,
976 Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-
977 binding sequencing experiments. Nucleic Acids Res. 2022 Jan 7;50(D1):D316–25.
978 61. Granja JM, Klemm S, McGinnis LM, Kathiria AS, Mezger A, Corces MR, et al. Single-cell
979 multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. Nat
980 Biotechnol. 2019 Dec 1;37(12):1458–65.
981 62. 10x Genomics. PBMC from a Healthy Donor - Granulocytes Removed Through Cell Sorting
982 (10k) [Internet]. 2021 [cited 2023 Sep 6]. Available from:
983 https://www.10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donor-
984 granulocytes-removed-through-cell-sorting-10-k-1-standard-2-0-0
985 63. Regner MJ, Wisniewska K, Garcia-Recio S, Thennavan A, Mendez-Giraldez R, Malladi VS, et al.
986 A multi-omic single-cell landscape of human gynecologic malignancies. Mol Cell.
987 2021;81(23):4924-4941.e10.
988 64. Kumegawa K, Saeki S, Takahashi Y, Yang L, Osako T, Nakadai T, et al. Chromatin profile-based

989 identification of a novel ER-positive breast cancer subgroup with reduced ER-responsive
990 element accessibility. Br J Cancer. 2023 Feb 1;128(7):1208–22.
991 65. Jin H, Liu Z. A benchmark for RNA-seq deconvolution analysis under dynamic testing
992 environments. Genome Biol. 2021 Apr 12;22:102.
993 66. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set
994 enrichment analysis: a knowledge-based approach for interpreting genome-wide expression
995 profiles. PNAS. 2005 Sep 30;102(43):15545–50.
996 67. Jiang Y, Hu Z, Jiang J, Zhu A, Zhang Y, Lynch AW, et al. scATAnno: Automated Cell Type
997 Annotation for single-cell ATAC Sequencing Data. bioRxiv. 2023 Jun 15;
998 68. Watt S, Vasquez L, Walter K, Mann AL, Kundu K, Chen L, et al. Genetic perturbation of PU.1
999 binding and chromatin looping at neutrophil enhancers associates with autoimmune disease.
1000 Nat Commun. 2021 Apr 16;12:2298.
1001 69. Geng D, Zheng L, Srivastava R, Asprodites N, Velasco-Gonzalez C, Davila E. When Toll-like
1002 receptor and T-cell receptor signals collide: a mechanism for enhanced CD8 T-cell effector
1003 function. Blood. 2010 Nov 4;116(18):3494–504.
1004 70. Javaid N, Choi S. Toll-like Receptors from the Perspective of Cancer Treatment. Cancers
1005 (Basel). 2020 Jan 27;12(2):297.
1006 71. Kalafati L, Kourtzelis I, Schulte-Schrepping J, Li X, Hatzioannou A, Grinenko T, et al. Innate
1007 Immune Training of Granulopoiesis Promotes Anti-tumor Activity. Cell. 2020 Oct
1008 29;183(3):771-785.e12.
1009 72. Sanseviero E. NK Cell-Fc Receptors Advance Tumor Immunotherapy. J Clin Med. 2019 Oct
1010 12;8(10):1667.
1011 73. Bonnema JD, Karnitz LM, Schoon RA, Abraham RT, Leibson PJ. Fc receptor stimulation of
1012 phosphatidylinositol 3-kinase in natural killer cells is associated with protein kinase C-
1013 independent granule release and cell-mediated cytotoxicity. J Exp Med. 1994 Oct
1014 1;180(4):1427–35.

1015 74. Smith JP, Corces MR, Xu J, Reuter VP, Chang HY, Sheffield NC. PEPATAC: an optimized pipeline
1016 for ATAC-seq data analysis with serial alignments. NAR Genom Bioinform. 2021 Dec;3(4).
1017 75. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis
1018 of ChIP-Seq (MACS). Genome Biol. 2008 Sep 17;9(9):R137.
1019 76. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning
1020 sequence reads to genomic features. Bioinformatics. 2014 Nov 13;30(7):923–30.
1021 77. Mayer M. R package ‘ splitTools ’: Tools for Data Splitting. R package version 1.0.1.
1022 https://cran.r-project.org/web/packages/splitTools/index.html. 2023.
1023 78. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression
1024 analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr
1025 20;43(7):e47.
1026 79. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential
1027 expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139–40.
1028 80. Law CW, Chen Y, Shi W, Smyth GK. Voom: Precision weights unlock linear model analysis tools
1029 for RNA-seq read counts. Genome Biol. 2014 Feb 3;15:R29.
1030 81. Lau JW, Lehnert E, Sethi A, Malhotra R, Kaushik G, Onder Z, et al. The cancer genomics cloud:
1031 Collaborative, reproducible, and democratized - A new paradigm in large-scale computational
1032 research. Cancer Res. 2017 Nov 1;77(21):e3–6.
1033 82. Kuhn M. Building Predictive Models in R Using the caret Package. J Stat Softw. 2008 Nov
1034 10;28(5):1–26.
1035 83. Risso D, Schwartz K, Sherlock G, Dudoit S. GC-Content Normalization for RNA-Seq Data. BMC
1036 Bioinformatics. 2011 Dec 17;12:480.
1037 84. Ucar D, Márquez EJ, Chung CH, Marches R, Rossi RJ, Uyar A, et al. The chromatin accessibility
1038 signature of human immune aging stems from CD8+ T cells. J Exp Med. 2017 Sep
1039 13;214(10):3123–44.
1040 85. Carvalho K, Rebboah E, Jansen C, Williams K, Dowey A, McGill C, et al. Uncovering the Gene

1041 Regulatory Networks Underlying Macrophage Polarization Through Comparative Analysis of
1042 Bulk and Single-Cell Data. bioRxiv. 2021 Jan 21;
1043 86. Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, Chang HY, et al. ArchR is a scalable
1044 software package for integrative single-cell chromatin accessibility analysis. Nat Genet. 2021
1045 Feb 25;53:403–11.
1046 87. Kumegawa K. ATAC-seq data of 42 BC samples as SummarizedExperiment object with count
1047 matrix, normalized count matrix, peak info, and clinical info [Internet]. 2023. Available from:
1048 https://doi.org/10.6084/m9.figshare.21992609.v1
1049

1050 **List of abbreviations:**

1051 **ATAC:** Assay for Transposase-Accessible chromatin

1052 **CBP(s):** chromatin binding protein(s)

1053 **ChIP-seq:** chromatin immunoprecipitation followed by sequencing

1054 **DC:** dendritic cells

1055 **NK:** natural killer cells

1056 **PCA:** principal component analysis

1057 **RMSE:** root mean squared error

1058 **TCGA :** The Cancer Genome Atlas

1059 **TF(s):** transcription factor(s)

1060 **TSS:** transcription start site

1061 **UMAP:** Uniform Manifold Approximation

1062

Main Figures.

Figure 1

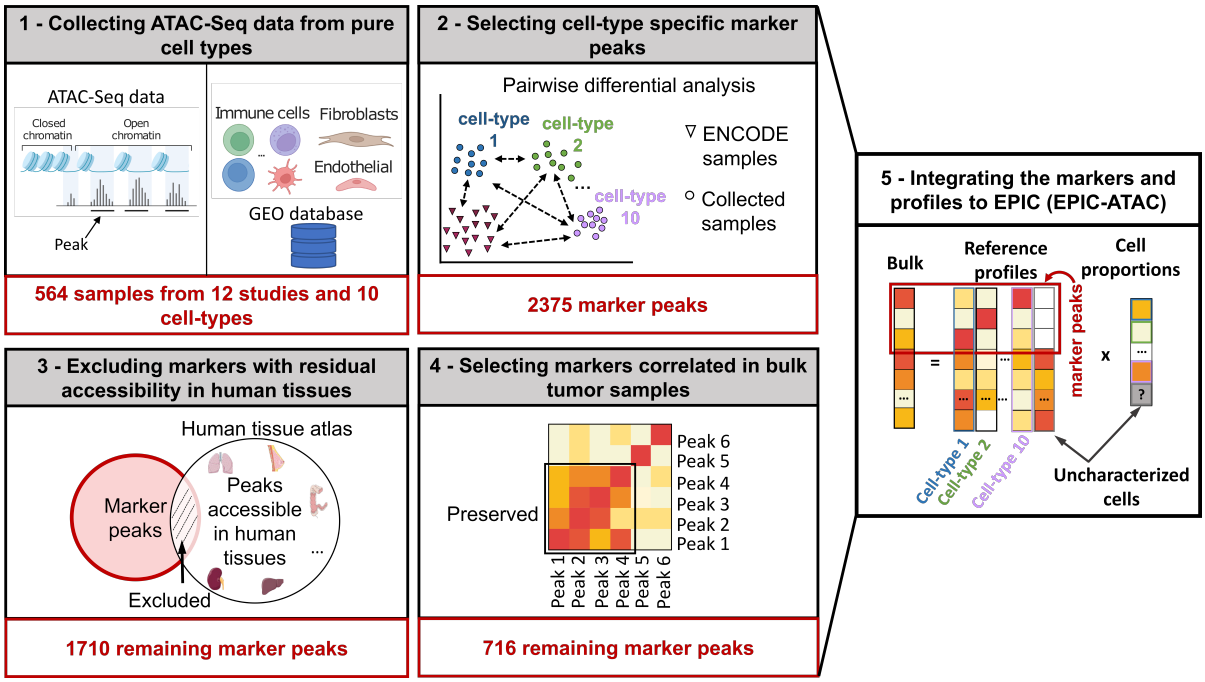


Figure 2

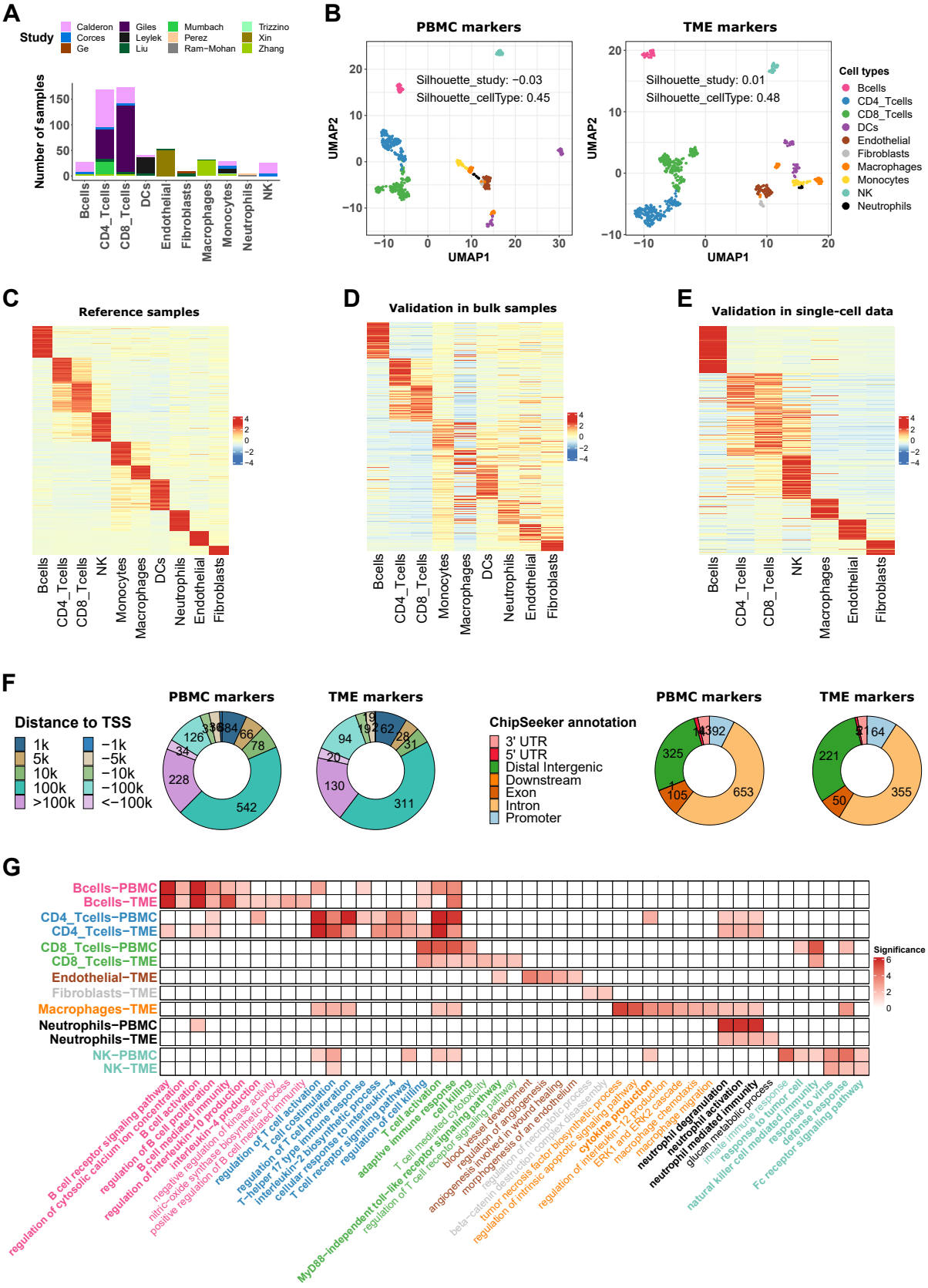

A
Study
Calderon
Corces
Ge
Giles
Leylek
Liu
Mumbach
Perez
Ram−Mohan
Trizzino
Xin
Zhang
Number of samples
Bcells
CD4_Tcells
CD8_Tcells
DCs
Endothelial
Fibroblasts
Macrophages
Monocytes
Neutrophils
NK
B
PBMC markers
Silhouette_study: −0.03
Silhouette_cellType: 0.45
TME markers
Silhouette_study: 0.01
Silhouette_cellType: 0.48
UMAP1
UMAP2
Cell types
C
Reference samples
D
Validation in bulk samples
E
Validation in single-cell data
F
Distance to TSS
1k
5k
10k
100k
>100k
−1k
−5k
−10k
−100k
<−100k
PBMC markers
TME markers
ChipSeeker annotation
3' UTR
5' UTR
Distal Intergenic
Downstream
Exon
Intron
Promoter
G
Significance

Figure 3

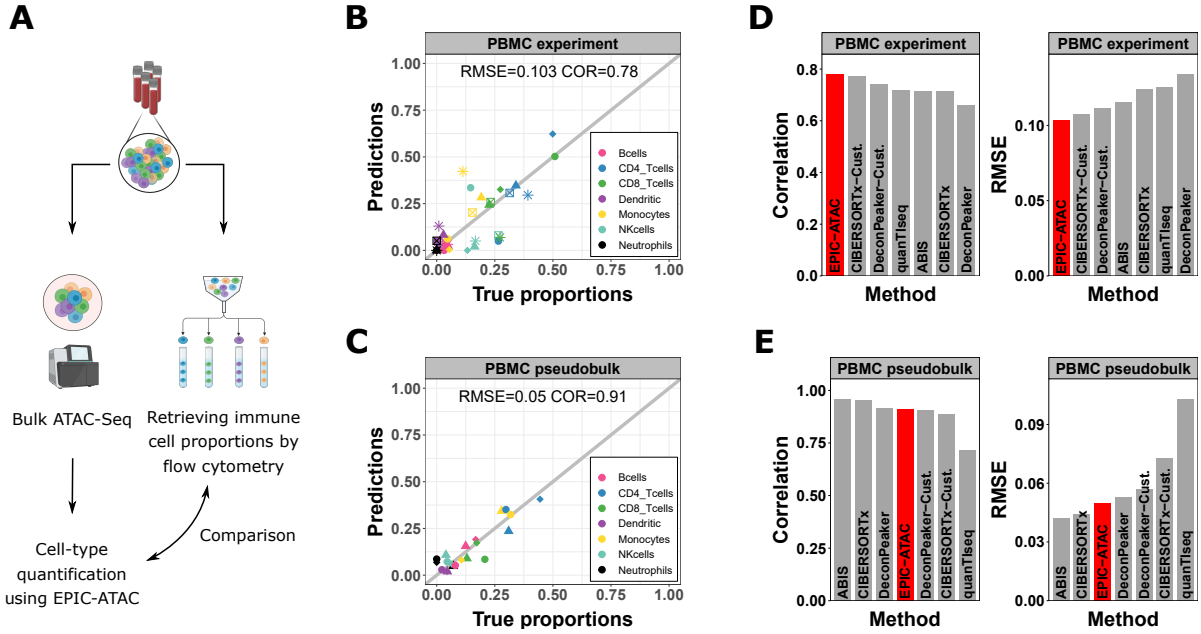

A
Bulk ATAC-Seq
Retrieving immune cell proportions by flow cytometry
Comparison
Cell-type quantification using EPIC-ATAC
B
PBMC experiment
RMSE=0.103 COR=0.78
Predictions
True proportions
Bcells
CD4_Tcells
CD8_Tcells
Dendritic
Monocytes
NKcells
Neutrophils
C
PBMC pseudobulk
RMSE=0.05 COR=0.91
Predictions
True proportions
D
PBMC experiment
Correlation
RMSE
Method
EPIC−ATAC
CIBERSORTx−Cust.
DeconPeaker−Cust.
quanTIseq
ABIS
CIBERSORTx
DeconPeaker
E
PBMC pseudobulk
Correlation
RMSE
Method
ABIS
CIBERSORTx
DeconPeaker
EPIC−ATAC
DeconPeaker−Cust.
CIBERSORTx−Cust.
quanTIseq

Figure 4

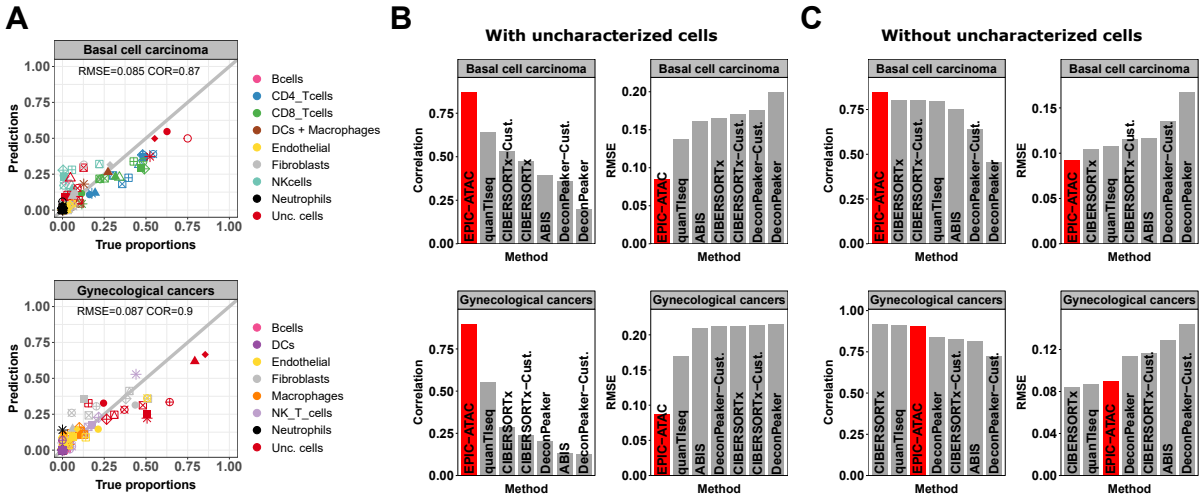

A
Basal cell carcinoma
RMSE=0.085 COR=0.87
Predictions
True proportions
Bcells
CD4_Tcells
CD8_Tcells
DCs + Macrophages
Endothelial
Fibroblasts
NKcells
Neutrophils
Unc. cells
Gynecological cancers
RMSE=0.087 COR=0.9
Bcells
DCs
Endothelial
Fibroblasts
Macrophages
NK_T_cells
Neutrophils
Unc. cells
B
With uncharacterized cells
Correlation
RMSE
Method
EPIC−ATAC
quanTiseq
CIBERSORTx
CIBERSORTx−Cust.
DeconPeaker
ABIS
DeconPeaker−Cust.
C
Without uncharacterized cells

Figure 5

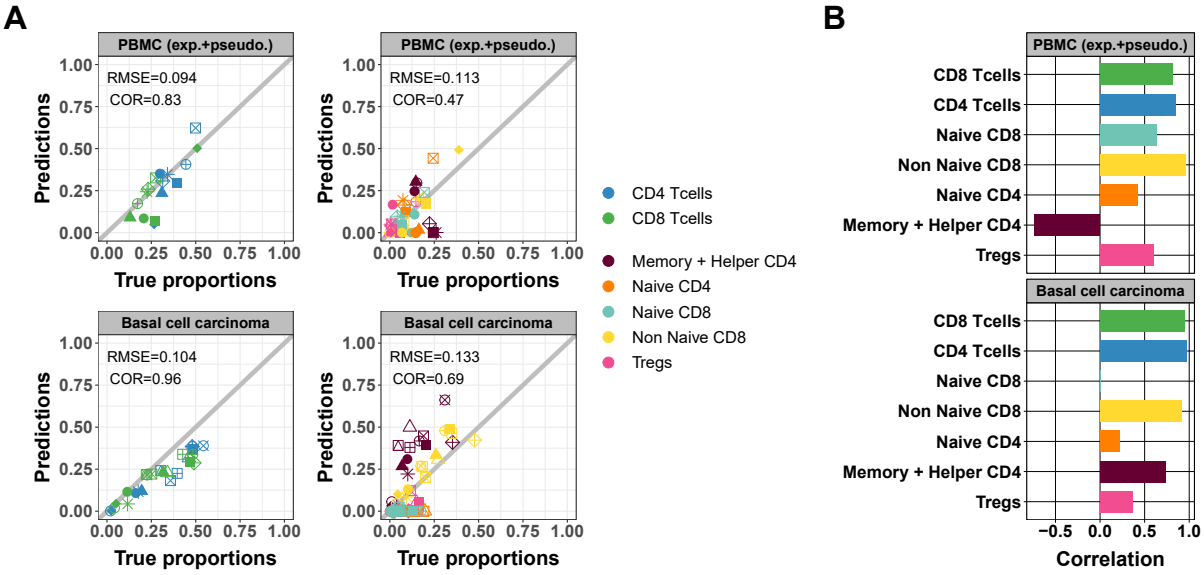

A
PBMC (exp.+pseudo.)
RMSE=0.094
COR=0.83
PBMC (exp.+pseudo.)
RMSE=0.113
COR=0.47
Basal cell carcinoma
RMSE=0.104
COR=0.96
Basal cell carcinoma
RMSE=0.133
COR=0.69
Predictions
True proportions
0.00
0.25
0.50
0.75
1.00
CD4 Tcells
CD8 Tcells
Memory + Helper CD4
Naive CD4
Naive CD8
Non Naive CD8
Tregs
B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

Figure 6

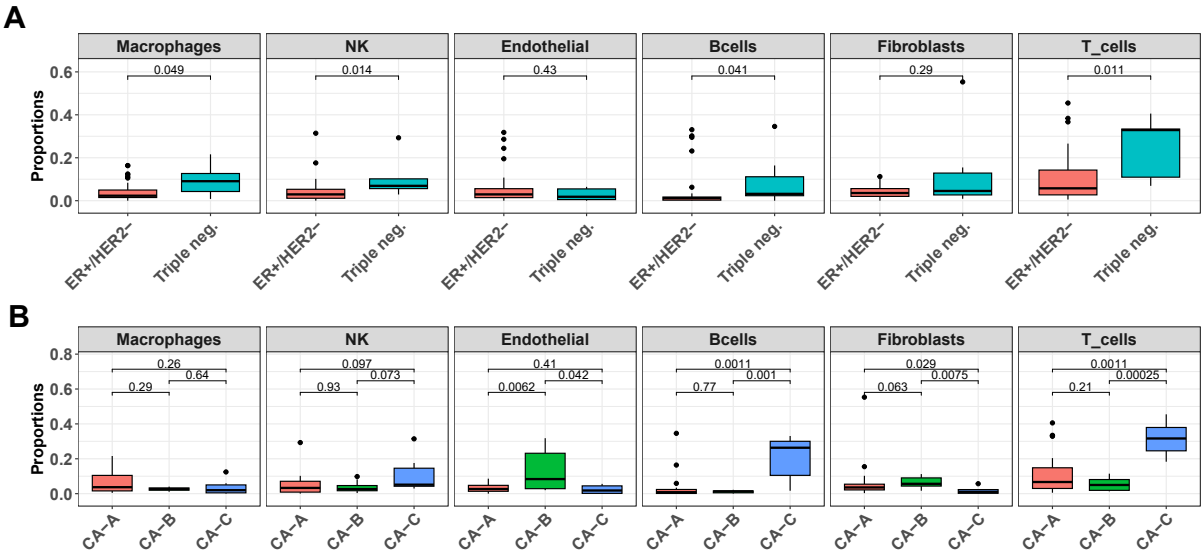
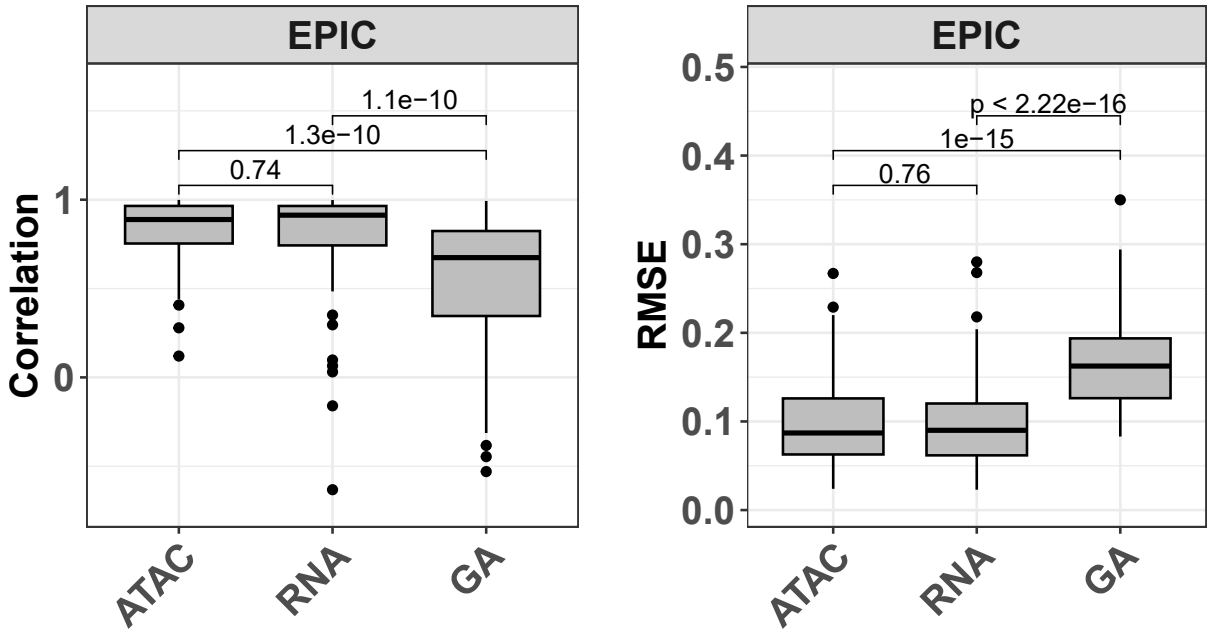

A
Macrophages
NK
Endothelial
Bcells
Fibroblasts
T_cells
Proportions
0.0
0.2
0.4
0.6
0.049
0.014
0.43
0.041
0.29
0.011
ER+/HER2−
Triple neg.
B
0.8
0.26
0.64
0.29
0.097
0.073
0.93
0.41
0.042
0.0062
0.0011
0.001
0.77
0.029
0.0075
0.063
0.0011
0.00025
0.21
CA−A
CA−B
CA−C

Figure 7

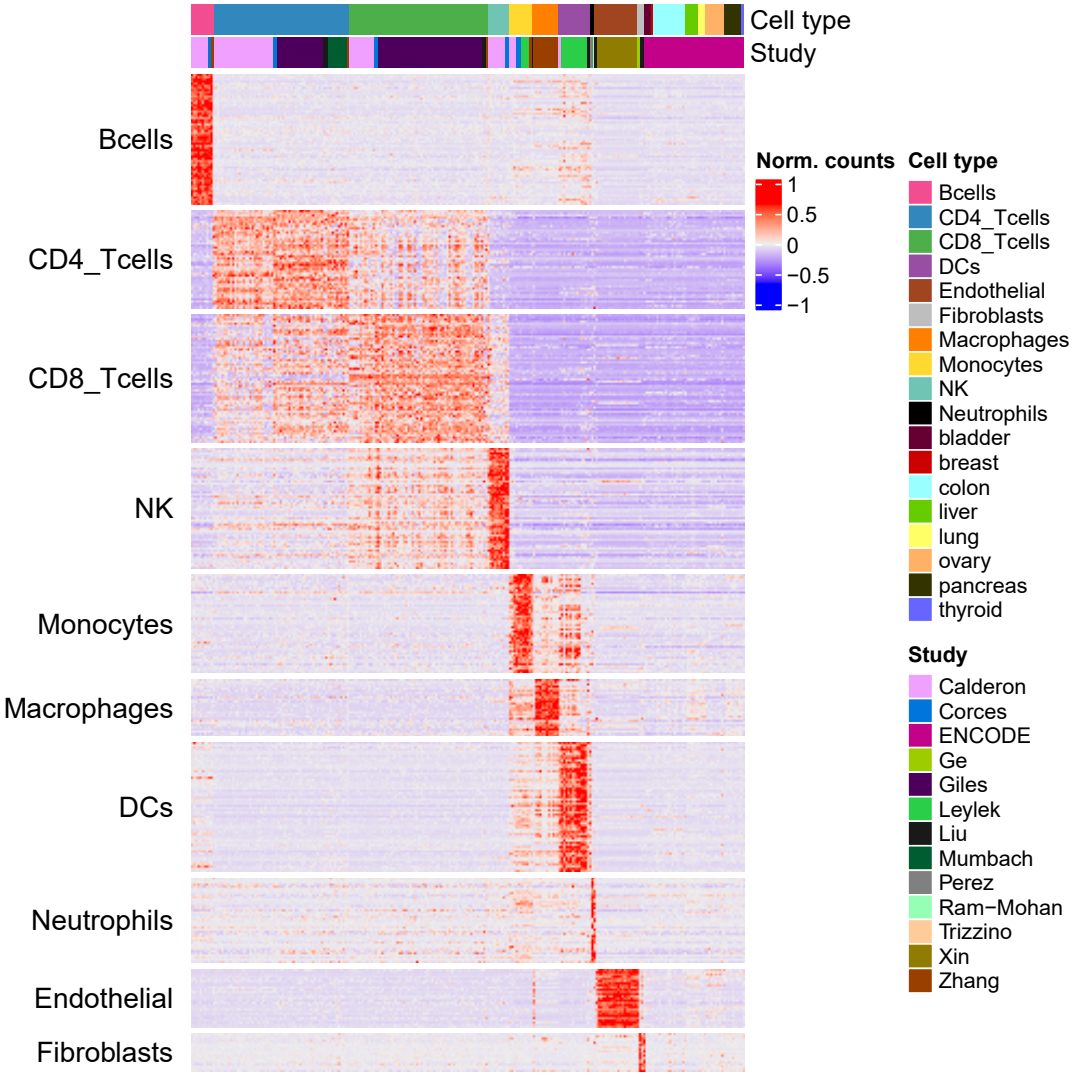

EPIC
Correlation
1
0
0.74
1.3e−10
1.1e−10
ATAC
RNA
GA
EPIC
RMSE
0.5
0.4
0.3
0.2
0.1
0.0
0.76
1e−15
p < 2.22e−16
ATAC
RNA
GA

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

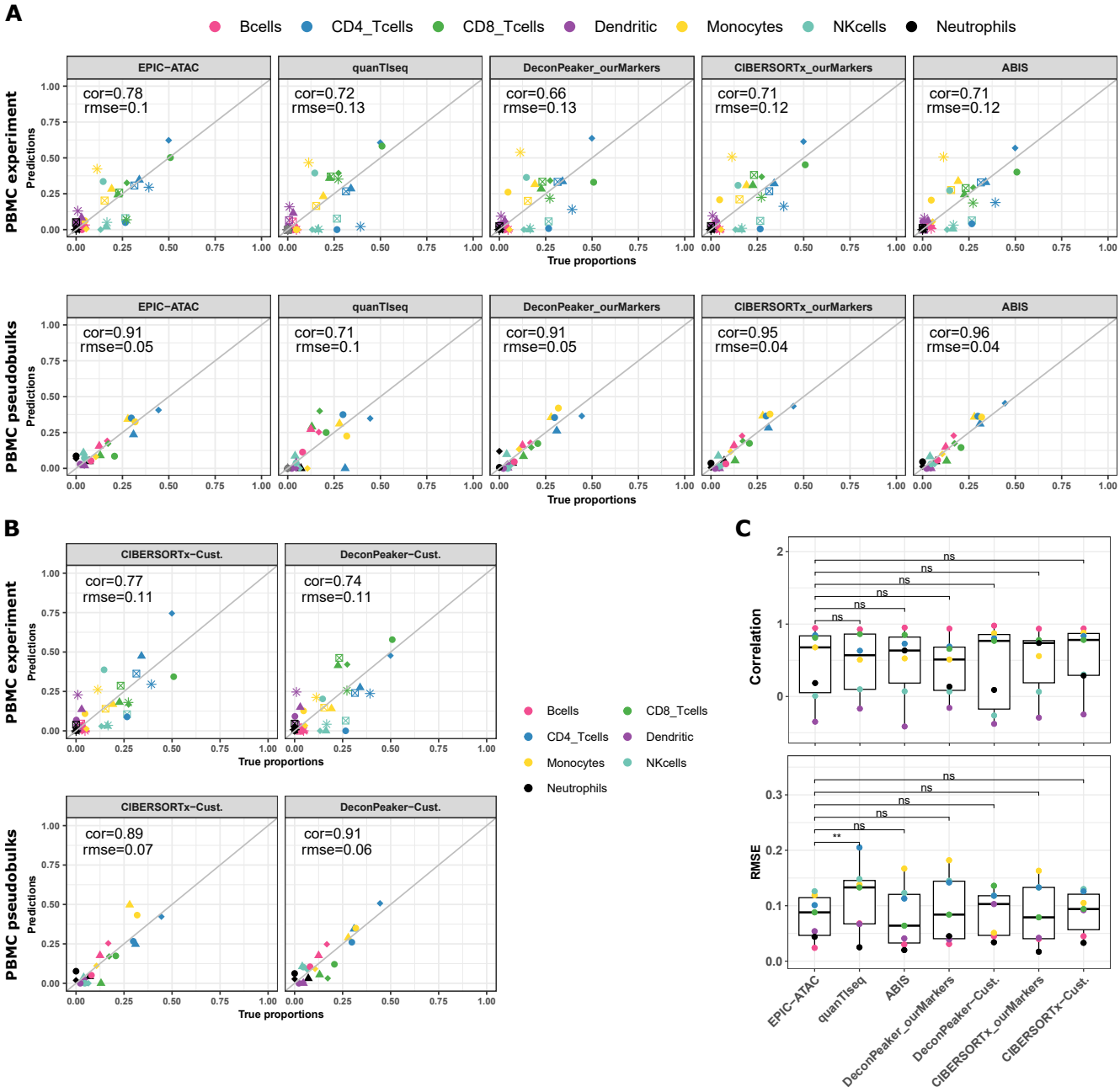
Aurélie AG Gabriel[1,2,3,4], Julien Racle[1,2,3,4], Maryline Falquet[3,5,6,7], Camilla Jandus[3,5,6,7], David Gfeller[1,2,3,4,*]

1 - Department of Oncology, Ludwig Institute for Cancer Research, University of Lausanne, Lausanne, Switzerland
2 - Agora Cancer Research Centre, 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

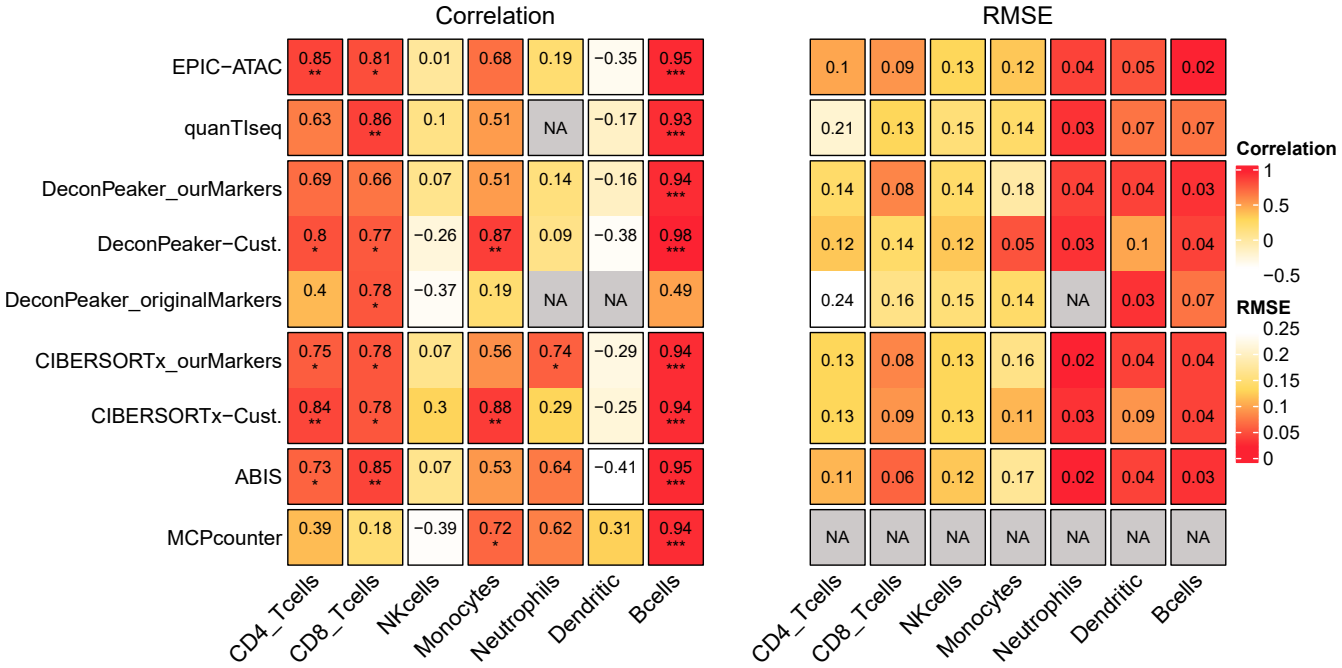
* Corresponding author: david.gfeller@unil.ch



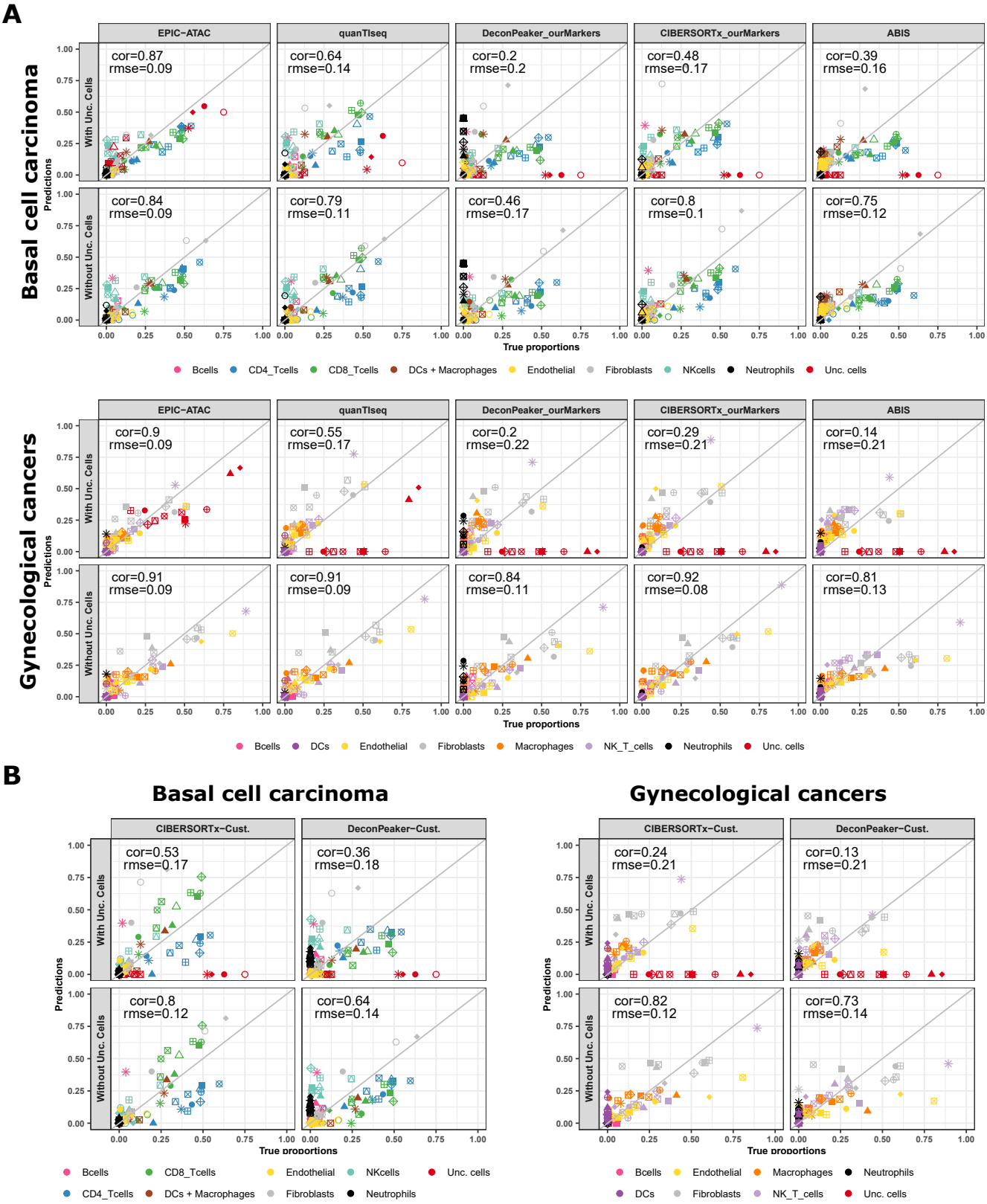
Supplementary Figure 1: Chromatin accessibility signal of each marker peak (rows) in each reference sample and in the ENCODE samples from diverse tissues (columns).



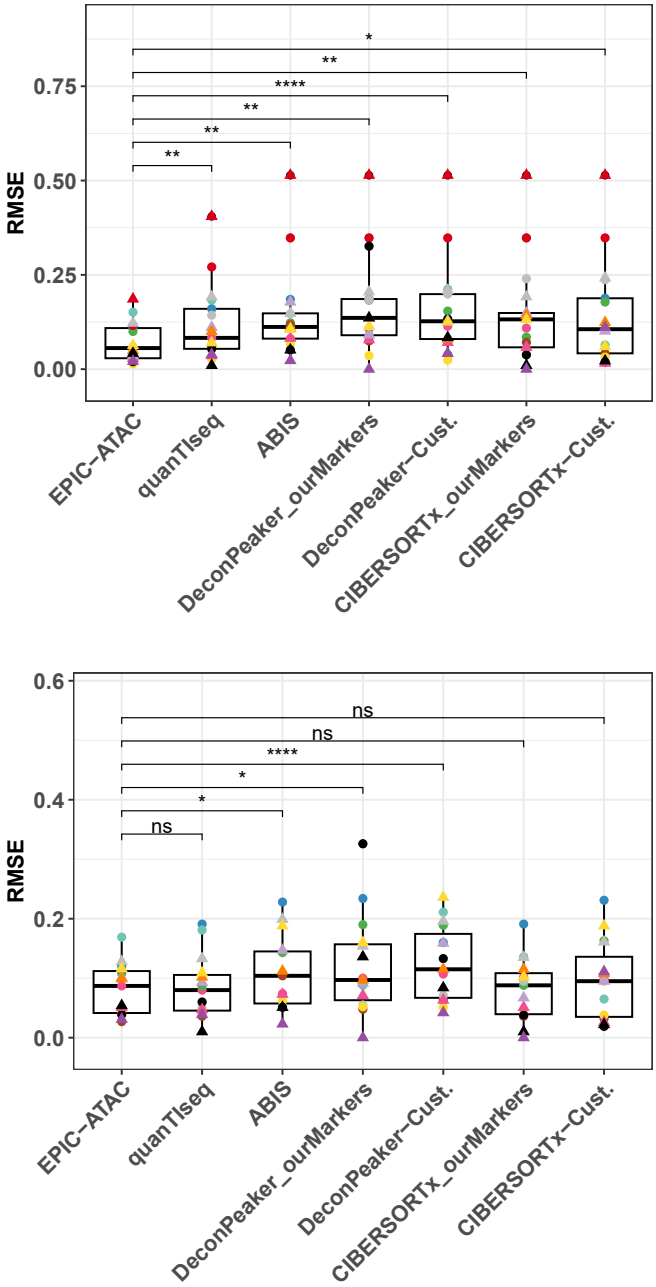
Supplementary Figure 2: A) Comparison of the cell-type proportions estimated by the different deconvolution methods included in the benchmark (y axis) and the true proportions (x axis) in the PBMC experiment samples (top panel) and in the PBMC pseudobulk samples (bottom panel). **B)** Comparison of the cell-type proportions estimated by CIBERSORTx and DeconPeaker using reference profiles built by each tools 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 and shapes to the PBMC dataset. Note that for quanTIseq, gray dots correspond to uncharacterized cells since the prediction of an uncharacterized cell type can not be turned off using the quanTIseq function from the quantiseqr R package. P-values resulting from a paired wilcoxon test are represented for each pair of comparison between tools. Significance levels are represented as follow: *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.



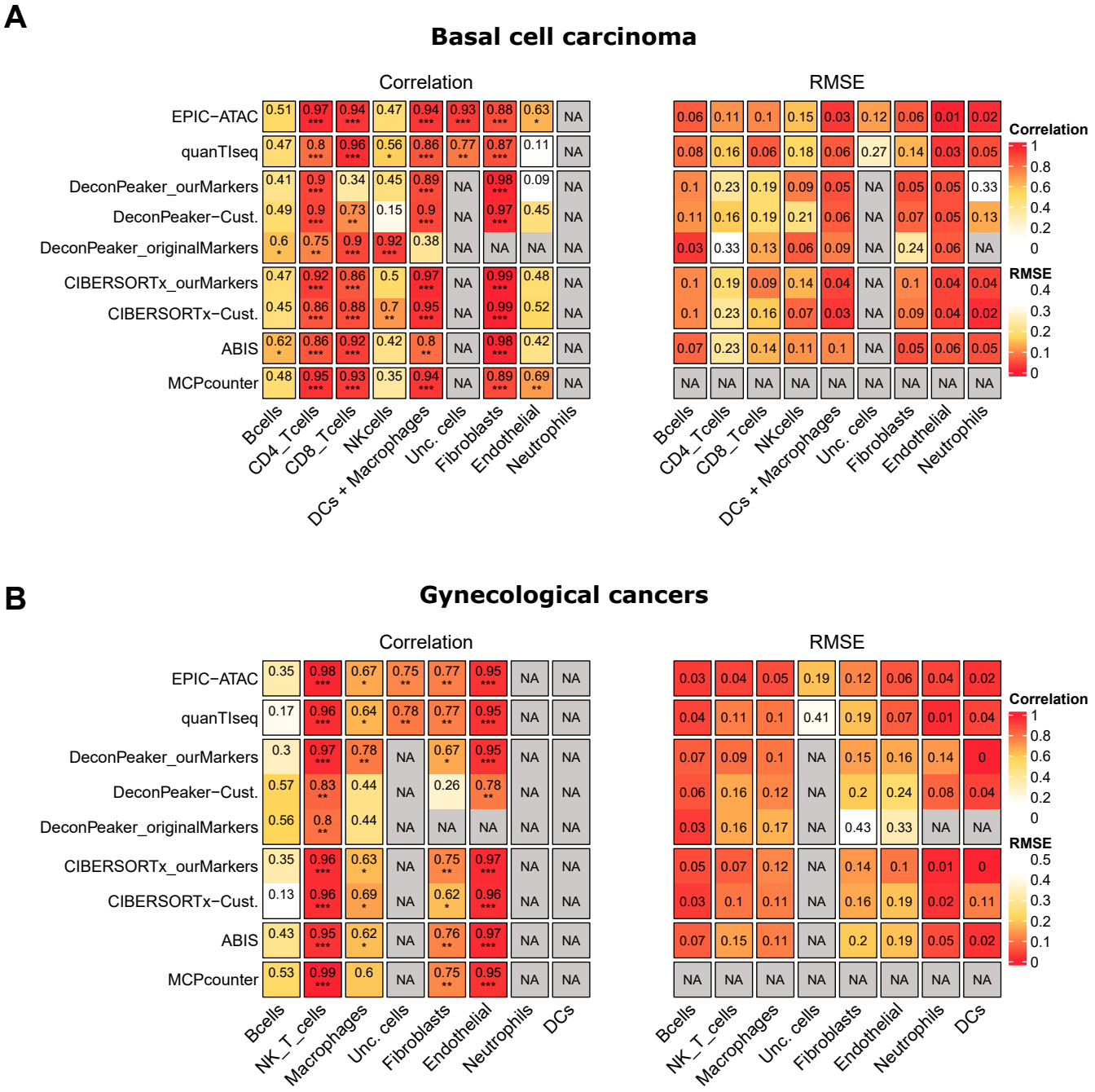
Supplementary Figure 3: Pearson's correlation (left panel) and RMSE values (right panel) associated to the comparison of the cell-type proportions estimated by each deconvolution tool (rows) included in the benchmark and the true cell-type proportions. The comparisons were made in each cell-type (columns) separately. Significance level of the correlations are represented as follow: *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.



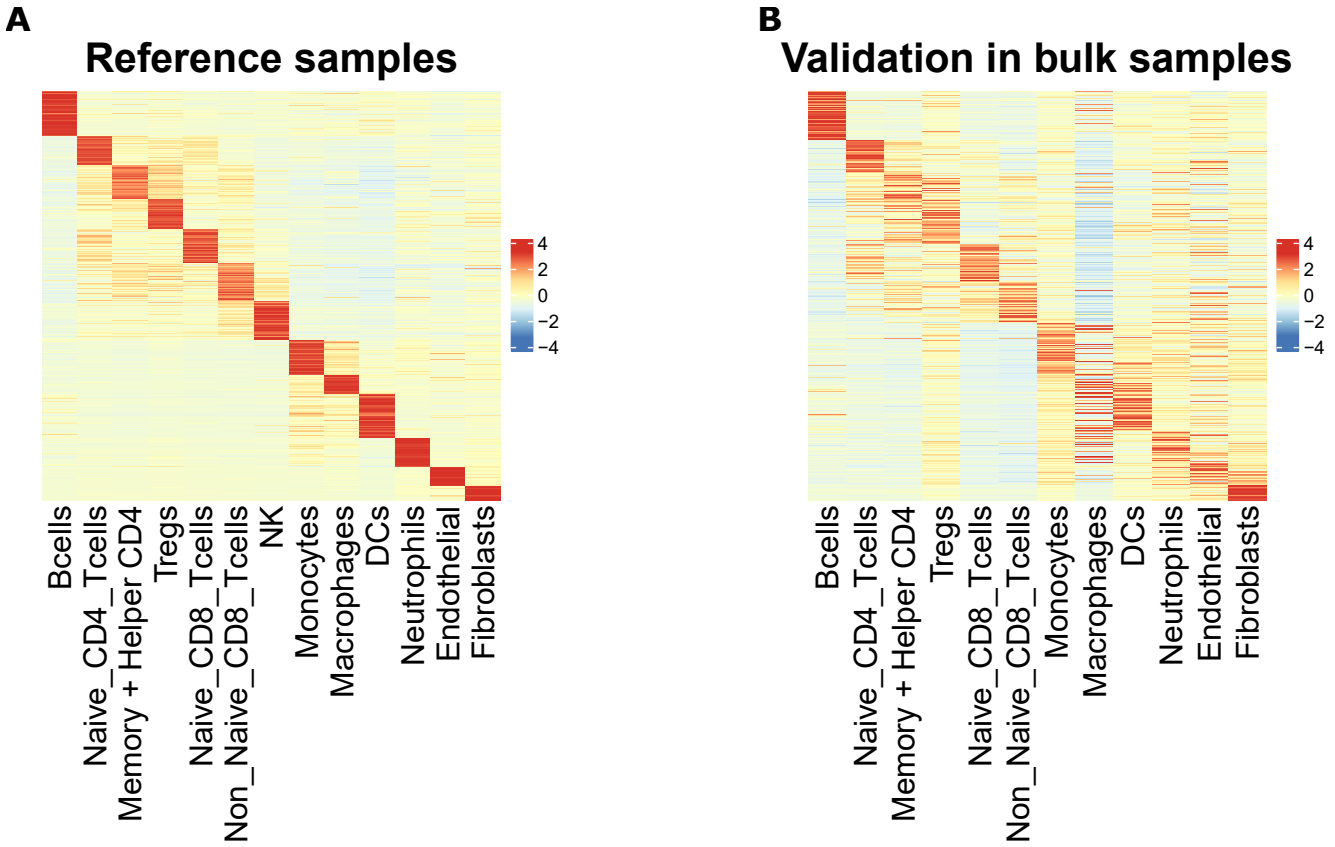
Supplementary Figure 4: A) Comparison of the cell-type proportions estimated by the different deconvolution methods using our reference profiles and the true proportion in the basal cell carcinoma samples (top panel) and in the gynecological cancer samples (bottom panel). In each panel, the top row includes the uncharacterized cells while the bottom row excludes them. **B)** Comparison of the cell-type proportions estimated by CIBERSORTx and DeconPeaker using reference profiles built by each tools using our collection of pure ATAC-Seq samples and the true proportion in the basal cell carcinoma samples (left panel) and in the gynecological cancer samples (right panel). In each panel, the top row includes the uncharacterized cells while the bottom row excludes them.



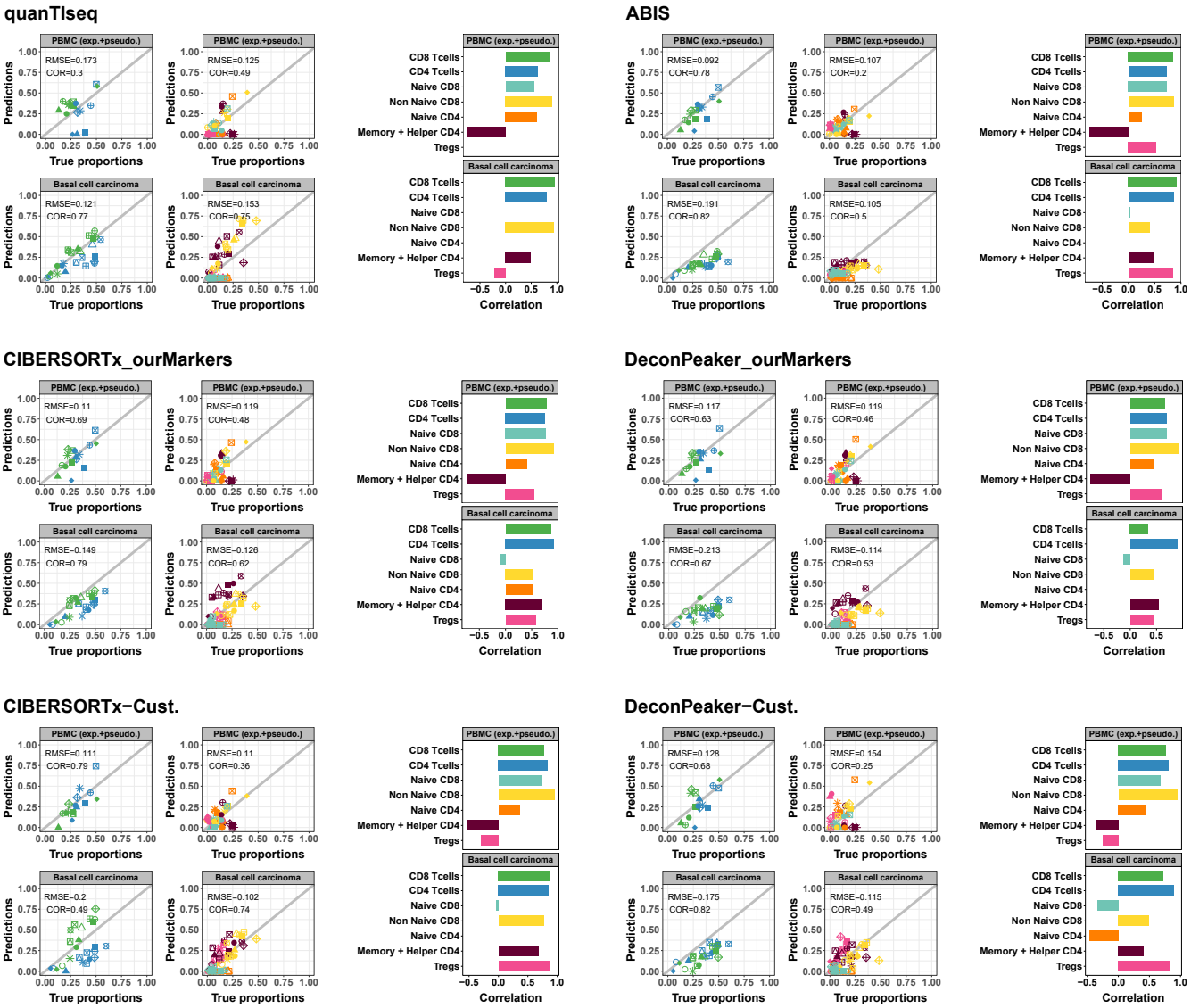
Supplementary Figure 5: Top panel: RMSE values associated to the comparison of the true proportions and the cell-type proportions estimated by each deconvolution tool in each cell type. For each cell type, all samples from each cancer pseudobulk datasets were considered,*i.e.*, colors correspond to cell 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 follow: *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. Bottom panel: Considering that not all deconvolution tools are able to predict the proportion of uncharacterized cells, the same analyis as in the top panel was performed without considering the uncharacterized cells predictions.



Supplementary Figure 6: 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** and **B** are associated to the benchmark performed on the basal cell carcinoma dataset and the gynecological cancer dataset respectively. Significance level of the correlations are represented as follow: *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.



Supplementary Figure 7: **A** Scaled averaged chromatin accessibility of each cell-type specific marker peak in each cell-type obtained based on the ATAC-Seq reference samples used to identify the marker peaks. **B** Scaled averaged chromatin accessibility of each cell-type specific marker peak in each cell-type from the sorted samples excluded from the reference samples (See Material and Methods).



Supplementary Figure 8: 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).