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

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

Affiliations:

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

[2] Agora Cancer Research Centre, Lausanne, Switzerland

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

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

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

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

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

* Corresponding author: david.gfeller@unil.ch

Abstract

Assay for Transposase-Accessible Chromatin sequencing (ATAC-Seq) is a widely used technique to explore gene regulatory mechanisms. For most ATAC-Seq data from healthy and diseased tissues such as tumors, chromatin accessibility measurement represents a mixed signal from multiple cell types. In this work, we derive reliable chromatin accessibility marker peaks and reference profiles for all major cancer-relevant cell types. We then capitalize on the EPIC deconvolution framework (Racle et al. 2017) previously shown to accurately predict cell-type composition in tumor bulk RNA-Seq data and integrate our markers and reference profiles to EPIC to quantify cell-type heterogeneity in bulk

ATAC-Seq data. Our EPIC-ATAC tool accurately predicts non-malignant and malignant cell fractions in tumor samples. When applied to a breast cancer cohort, EPIC-ATAC accurately infers the immune contexture of the main breast cancer subtypes.

Introduction

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

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

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

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

102 data from neutrophils due to extracellular traps formation). Also, these markers have not been
103 curated to fulfill the requirements of tools such as EPIC to quantify uncharacterized cells (*i.e.*,
104 markers of a cell-type should not be accessible in other human tissues).
105
106 In this study, we collected ATAC-Seq data from pure cell types to identify cell-type specific
107 marker peaks and to build reference profiles from most major non-malignant cell types typically
108 observed in tumors. These data were integrated in the EPIC (Racle et al. 2017) framework to perform
109 bulk ATAC-Seq samples deconvolution (Figure 1). Applied on peripheral blood mononuclear cells
110 (PBMCs) and tumor samples, the EPIC-ATAC framework showed accurate predictions of the
111 proportions of non-malignant and malignant cells with similar or higher performances than other
112 existing tools.
113
114

115 ## Results

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

118 A key determinant for accurate predictions of cell-type proportions by most deconvolution tools is
119 the availability of reliable cell-type specific markers and reference profiles. To identify robust
120 chromatin accessibility marker peaks of cancer relevant cell types, we collected 564 samples of
121 sorted cell populations from twelve studies including eight immune cell types (B cells (Calderon et al.
122 2019; Corces et al. 2016; P. Zhang et al. 2022), CD4+ T cells (Corces et al. 2016; Liu et al. 2020; P.
123 Zhang et al. 2022; Mumbach et al. 2017; Giles et al. 2022), CD8+ T cells (Calderon et al. 2019; Corces
124 et al. 2016; Liu et al. 2020; P. Zhang et al. 2022; Giles et al. 2022), natural killer (NK) cells (Calderon et
125 al. 2019; Corces et al. 2016), dendritic cells (DCs) (Calderon et al. 2019; Leylek et al. 2020; Liu et al.
126 2020), macrophages (Liu et al. 2020; P. Zhang et al. 2022), monocytes (Calderon et al. 2019; Corces et
127 al. 2016; Leylek et al. 2020; P. Zhang et al. 2022; Trizzino et al. 2021) and neutrophils (Ram-Mohan et

128 al. 2021; Perez et al. 2020), as well as fibroblasts (Ge et al. 2021; Liu et al. 2020) and endothelial (Liu
129 et al. 2020; Xin et al. 2020) cells (Figure 1 box 1, Figure 2A, Supplementary Table 1). To limit batch
130 effects, the collected samples were homogeneously processed from read alignment to peak calling.
131 For each cell type, we derived a set of stable peaks, *i.e.,* peaks observed across samples and studies
132 (see Materials and Methods).
133 These peaks were then used to perform pairwise differential analysis to identify marker peaks for
134 each cell type (Figure 1, box 2). To ensure that the cell-type specific marker peaks are not accessible
135 in other human tissues, we included in the differential analysis ATAC-Seq samples from diverse
136 human tissues from the ENCODE data (The ENCODE Project Consortium et al. 2020; Rozowsky et al.
137 2023) (Supplementary Figure 1). To select a sufficient number of peaks prior to peak filtering, the top
138 200 peaks recurrently differentially accessible across all cell-type pairs were selected as cell-type
139 specific markers (see Materials and Methods). Using the human atlas study (K. Zhang et al. 2021),
140 markers with potential residual accessibility in human tissues were then filtered out (Figure 1, box 3,
141 see Materials and Methods). The resulting marker peaks specific to the immune cell types were
142 considered for the deconvolution of PBMC samples (PBMC markers). For tumor bulk sample
143 deconvolution, the list of markers was further refined based on the correlation patterns of the
144 markers in tumor bulk samples from diverse cancer types from The Cancer Genome Atlas (TCGA)
145 (Corces et al. 2018) (Figure 1, box 4, see the Material and methods). The latter filtering ensures the
146 relevance of the markers in the TME context since cell-type specific TME markers are expected to be
147 correlated in tumor bulk ATAC-Seq measurements (Qiu et al. 2021). 716 markers of immune,
148 fibroblasts and endothelial cell types remained after the later filtering and were considered for the
149 deconvolution of bulk tumor samples (TME markers).
150 To assess the quality and reproducibility of these markers, we performed principal component
151 analysis (PCA) based on each set of marker peaks. Computing silhouette coefficients based on the
152 cell-type classification and on the study of origin showed that samples clustered by cell type and not
153 by study of origin (averaged silhouette coefficients above 0.45 for cell type and around 0 for study of

154 origin). Two-dimensional UMAP representations of the samples confirmed this observation (Figure
155 2B). These results indicate limited remaining batch effects after data processing and marker
156 selection.
157 We then used the collected samples to generate chromatin accessibility profiles by computing the
158 average of the normalized counts for each peak in each cell type as well as peak variability in each
159 cell type (Racle et al. 2017) (see Material and methods). Figure 2C represents the average chromatin
160 accessibility of each marker peak in each cell type of the reference dataset and highlights, as
161 expected, the cell-type specificity of the selected markers (see also Supplementary Tables 2 and 3),
162 which was confirmed in independent ATAC-Seq data from sorted cells and single-cell ATAC-Seq
163 samples from blood and diverse human tissues (Figure 2D and 2E, see Materials and methods).
164

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

166 To characterize the different marker peaks, we annotated them using ChiPSeeker (Yu, Wang, and He
167 2015). We observed that most of the markers are in distal and intergenic regions (Figure 2F), which is
168 expected considering the large proportion of distal regions in the human genome and the fact that
169 such regions have been previously described as highly cell-type specific (Corces et al. 2016). We also
170 noticed that 7% of the PBMC and TME marker peaks are in promoter regions in contrast to 4% when
171 considering matched genomic regions randomly selected in the set of peaks identified prior to the
172 differential analysis (see Material and methods), which suggest enrichment in our marker peaks for
173 important regulatory regions.
174 To assess the biological relevance of the marker peaks, we associated each marker peak to its
175 nearest gene using ChIP-Enrich based on the "nearest transcription start site (TSS)" locus definition
176 (Welch et al. 2014) (Supplementary Tables 4 and 5). Nearest genes reported as known marker genes
177 in public databases of gene markers (*i.e.,* PanglaoDB (Franzén, Gan, and Björkegren 2019) and
178 CellMarker (Hu et al. 2023)) are listed in Table 1.

179 In each set of cell-type specific peaks, we observed an overrepresentation of chromatin binding
180 proteins (CBPs) reported in the JASPAR2022 database (Castro-Mondragon et al. 2022) (using Signac
181 (Stuart et al. 2021) and MonaLisa (Machlab et al. 2022) for assessing the overrepresentation) and the
182 ReMap catalog (Hammal et al. 2022) (using RemapEnrich, see Material and Methods).
183 Overrepresented CBPs also reported as known marker genes in the PanglaoDB and CellMarker
184 databases are listed in Table 1. Detailed peaks annotations are summarized in Supplementary Tables
185 4 and 5.
186 Based on the "nearest TSS" annotation, we tested, using ChIP-Enrich (Welch et al. 2014), whether
187 each set of cell-type specific marker peaks was enriched for regions linked to specific biological
188 pathways (GO pathways). Figure 2G highlights a subset of the enriched pathways that are consistent
189 with prior knowledge on each cell type. Some of these pathways are known to be characteristic of
190 immune responses to inflammatory or tumoral environments. The complete list of enriched
191 pathways is listed in the Supplementary Tables 6 and 7. Overall, these analyses demonstrate that the
192 proposed cell-type specific marker peaks capture some of the known biological properties associated
193 to each cell type.

Cell type	Nearest genes	Enriched CPBs
Bcells	DHTKD1 LHPP WDFY4 ARID5B HHEX SIDT2 CD82 MS4A1 FCHSD2 USP8 RHCG ATF7IP2 CIITA GGA2 SNX29P2 C16orf74 CBFA2T3 CD79B BCL2 GNG7 CD22 FCER2 FCRL1 LY9 PTPRC LAPTM5 IGLL5 VPREB3 CENPM AFF3 SP100 INPP5D DTNB CD86 RFTN1 ST6GAL1 NGLY1 OSBPL10 TLR9 CD38 SMIM14 ARHGAP24 ADAM19 EBF1 BASP1 CD83 PLEKHG1 CCR6 CCND3 HDAC9 CDCA7L BLK MTSS1 LYN PLEKHF2 MOB3B PAX5	SPIB POU2F2 TCF4 EBF1 TCF3 NFKB1 STAT1 NFKB2 IKZF1 FOXO1 FOXP1 BCL6 POU2AF1 STAT3 BACH2 IKZF3 FLI1 TBX21 JUNB MITF NKX6-2 RBPJ
CD4_Tcells	IL2RA CD6 CD5 CD4 RORA PTPRC CTLA4 ICOS SLC9A9 FHIT TCF7 FYB1 ATXN1 CD40LG	TCF7 RUNX3 SOHLH2 IRF9 GATA3 TBX21 MAF STAT3 RORA BATF CREM
CD8_Tcells	MKI67 JAML MAML2 KLRD1 NELL2 LAG3 PPP1R13B PTPRC LYST CASP8 CD8A CD8B CD96 BTLA GZMA THEMIS ETV1	ETV1 FOXP3 TBX21 FOXP1 EOMES CREM IRF4 ZEB1 ARNT JUNB TCF7
NK	PRF1 ZBTB16 KLRD1 SPN CD226 SH2D1B CD247 IL2RB CXCR4 NMUR1 GNLY ZAP70 TXK	EOMES TBX21 NFIL3 FOS JUN
DCs	C12orf75 LYZ APP CD8A RIOX2 NFKB1 QDPR ABCG2 PRELID2 DST CD36 IDO2 PCMTD1	SPIB IRF8 MYB NR4A1 REL CUX2 FOXO1 ETV6 IRF5 BATF3 RUNX2
Neutrophils	TLE3 CA4 CYP4F3 CEACAM8 PGLYRP1 FPR1 CTSS ALPL PI3 MMP9 CXCR1 DRC1 ASPRV1 LTF MGAM SLC25A37	FOS

Monocytes	VENTX GLT1D1 CLEC4E CARS2 SLC24A4 C16orf74 FFAR2 STXBP2 NLRP3 CYRIA CMTM7 TGFBI DIAPH1 VCAN MCTP1 IFNGR1 STX11 CAPZA2 CD36 MTSS1 DENND3 ASAH1 TNFRSF10B BNIP3L NACC2 MAMDC2 FBP1	CEBPA CEBPD CEBPB CEBPE SPI1 VENTX JUND RXRA TCF7L2
Macrophages	CXCL12 PSAP P2RY6 SLCO2B1 CMKLR1 MMP19 LGMN CLEC10A C5AR1 FPR3 LILRB4 RGL1 SIGLEC1 MMP9 CD80	STAT1 SPI1 FOSL2 FOS SPIC
Endothelial	FAM107B ROBO4 FLI1 ACVRL1 FLT1 DOCK9 ABCC1 S1PR1 ELOVL1 PLPP3 ASAP2 SNRK ECSCR ARAP3 LAMA4 BMP6 SERPINE1 LAMB1 DOCK4 NOS3	ETV2 ELF1 FLI1 ELK3 FOSB ETS1 ERG GATA2 ZEB1 ETS2 FOXC1 SOX18
Fibroblasts	LOX CAV1 COL15A1	FOSL2 FOSB FLI1 HIF1A PBX1

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

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

The cell-type specific marker peaks and profiles derived from the reference samples were integrated to the EPIC deconvolution tool (Racle et al. 2017; Racle and Gfeller 2020). We will refer to this ATAC-Seq deconvolution framework as EPIC-ATAC.

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

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

The accuracy of the predictions obtained with EPIC-ATAC was then compared with the accuracy of other deconvolution approaches which could be used with our reference profiles and marker peaks

215 (Figure 3D-E). To this end, we considered both the DeconPeaker method (H. Li et al. 2020) originally
216 developed for bulk ATAC-Seq as well as several algorithms developed for bulk RNA-Seq (CIBERSORTx
217 (Newman et al. 2019), QuanTIseq (Finotello et al. 2019), ABIS (Monaco et al. 2019), and MCPcounter
218 (Becht et al. 2016)). To enable meaningful comparison across the cell types considered in this work
219 and use the method initially developed for bulk RNA-Seq deconvolution, the marker peaks and
220 profiles derived in this work were used in each of these methods. DeconPeaker and CIBERSORTx
221 include the option to define cell-type specific markers and profiles from a set of reference samples.
222 We thus fed our ATAC-Seq samples collection to both algorithms and used the resulting profiles and
223 marker peaks to perform bulk ATAC-Seq deconvolution. The resulting predictions are referred to as
224 DeconPeaker-Custom and CIBERSORTx-Custom.
225 Many tools displayed high correlation and low RMSE values, similar to those of EPIC-ATAC, and no
226 single tool consistently outperformed the others (Figure 3D-E, Supplementary Figure 2A-C). The fact
227 that our marker peaks and reference profiles could be used with EPIC-ATAC and other existing tools
228 demonstrates their broad applicability.
229 Predictions accuracies were also evaluated in each cell type separately. Since the number of samples
230 was low in each dataset, samples from both datasets were combined for this analysis. EPIC-ATAC
231 demonstrated good accuracies across cell types with RMSE values ranging from 0.02 for B cells to
232 0.13 for NK cells (Supplementary Figure 3). As expected, predictions with all tools were more
233 accurate for frequent cell types with well-characterized markers (*e.g.,* CD8/CD4 T cells, B cells)
234 compared to less frequent cell types (*e.g.,* NK cells, dendritic cells) (Supplementary Figure 2 and 3).
235 Note that MCPcounter is a marker-based method that derives cell-type specific scores which cannot
236 be compared between cell types. This method was thus only included in the benchmark considering
237 each cell type separately.
238
239 **EPIC-ATAC accurately predicts fractions of cancer and non-malignant cells in tumor**
240 **samples**

241 We evaluated the ability of the EPIC-ATAC framework to predict not only immune and stromal cells
242 proportions but also the proportion of cells for which reference profiles are not available (*i.e.,*
243 uncharacterized cells). For this purpose, we considered two previously published scATAC-Seq
244 datasets containing basal cell carcinoma and gynecological cancer samples (Satpathy et al. 2019;
245 Regner et al. 2021). We generated two pseudobulk datasets by averaging the chromatin accessibility
246 signal across all cells of each sample (see Material and methods). Applying EPIC-ATAC to both
247 datasets shows that this framework is able to simultaneously predict the proportions of both
248 uncharacterized cells and immune, stromal and vascular cells (Figure 4A). In these cancer samples,
249 the proportion of uncharacterized cells can be seen as a proxy of the proportion of cancer cells.
250 As for the PBMC datasets, we compared EPIC-ATAC performances to other existing deconvolution
251 tools. For both datasets, EPIC-ATAC led to the highest performances and was the only method to
252 accurately predict the proportion of uncharacterized cells (Figure 4B, Supplementary Figure 4 and 5).
253 Although quanTIseq also allows users to perform such predictions, the method resulted in lower
254 correlation and higher RMSE values when comparing the estimated and true proportions of the
255 uncharacterized cells (Figure 4B, Supplementary Figure 4).
256 In the EPIC-ATAC and quanTIseq frameworks, predictions correspond to absolute cell-type fraction,
257 *i.e.,* proportions of all cells present in the bulk, while the estimations obtained from the other tools
258 correspond to relative cell fractions, *i.e.,* proportions of cells present in the reference profiles
259 (CIBERSORTx, DeconPeaker) or to scores with arbitrary units (ABIS, MCPcounter). We thus conducted
260 a second benchmark excluding the predictions of uncharacterized cell fractions and rescaling both
261 estimations and true proportions to sum to 1 (see Material and methods). EPIC-ATAC outperformed
262 most of the other methods also when excluding the uncharacterized cells (Figure 4C, Supplementary
263 Figure 4 and 5).
264 Supplementary Figure 6 reports the performances of each tool when considering each cell type
265 separately. Overall, EPIC-ATAC showed comparable or higher correlation and lower RMSE values
266 when compared to the other deconvolution tools.

267

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

270 To explore the limitations of ATAC-Seq deconvolution, we next evaluated whether EPIC-ATAC could
271 predict the proportions of T-cell subtypes. To this end, we considered naive and non-naive CD8+ as
272 well as naïve, helper/memory and T regulatory CD4+ T cells. We redefined our list of cell-type specific
273 marker peaks and reference profiles including also these five T-cell subtypes (Supplementary Tables
274 8-9, Supplementary Figure 7A) and observed that the markers were conserved in external data
275 (Supplementary Figure 7B). The annotations of the markers associated to the T-cell subtypes are
276 available in Supplementary Tables 10-13.

277 We capitalized on the more detailed cell-type annotation of the PBMC datasets as well as the basal
278 cell carcinoma dataset to evaluate the EPIC-ATAC prediction of cell-subtype fractions using these
279 updated markers and profiles. Overall, the correlations observed between the predictions and true
280 proportions of T cells decreased when considering T-cell subtypes rather than CD4+ and CD8+ cell
281 types only (Figure 5A). In particular, low accuracies were obtained for helper/memory CD4+ and
282 naïve T-cell subtypes (Figure 5B). Similar results were obtained using other deconvolution tools
283 (Supplementary Figure 8).

284

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

287 We applied EPIC-ATAC to a breast cancer cohort of 42 breast ATAC-Seq samples including samples
288 from two breast cancer subtypes, *i.e.,* 35 oestrogen receptor (ER)-positive human epidermal growth
289 factor receptor 2 (HER2)-negative (ER+/HER2-) samples and 7 triple negative (TN) tumors (Kumegawa
290 et al. 2023). No cell sorting was performed in parallel to the chromatin accessibility sequencing. We
291 thus used EPIC-ATAC to estimate cell-type proportions. We observed a higher proportion of T cells, B
292 cells, NK cells and macrophages in the TN samples in comparison to ER+/HER2- samples (Figure 6A).

293 We then compared the cellular composition of ER+/HER2- subgroups identified in the original study
294 (clusters CA-A, CA-B and CA-C). A higher infiltration of T and B cells was observed in cluster CA-C and
295 higher proportions of endothelial cells and fibroblasts were observed in cluster CA-B (Figure 6B).
296 These predictions are consistent with the infiltration level estimations reported in the original
297 publication, although no differences in macrophages infiltration was observed between the
298 ER+/HER2- subgroups in our case (Kumegawa et al. 2023).
299

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

302 We finally compared the accuracy of EPIC when applied on ATAC-Seq data and on RNA-Seq data. For
303 this purpose, we used the 10X multiome PBMC dataset (10x Genomics 2021) which provides for each
304 cell both its chromatin accessibility profile and its gene expression profile and simulated 100
305 pseudobulks with diverse cellular compositions (see Material and methods). We used EPIC-ATAC to
306 perform ATAC-Seq based deconvolution on the chromatin accessibility levels of the peaks and the
307 original EPIC tool to perform standard RNA-seq deconvolution on the gene expression levels. ATAC-
308 Seq peaks can also be aggregated, based on peak distances to each gene, into gene activity (GA)
309 variables as proxy for gene expression. We thus applied the GA transformation to the 10x multiome
310 PBMC dataset and performed GA-based RNA deconvolution using the original EPIC tool (See Material
311 and methods).
312 Figure 7 shows that EPIC-ATAC performs similarly to the EPIC RNA-seq based deconvolution and
313 outperforms the GA-based RNA deconvolution. The lower performances of GA based RNA
314 deconvolution could be explained by the fact that GA features, by construction, do not perfectly
315 match the transcriptomic data.
316

317 **Discussion**

318 Bulk chromatin accessibility profiling of biological tissues like tumors represents a reliable and
319 affordable technology to map the activity of gene regulatory elements across multiple samples in
320 different conditions. Here, we collected ATAC-Seq data from pure cell populations covering major
321 immune and non-immune cancer-relevant cell types from diseased, stimulated and healthy samples.
322 This enabled us to identify reliable cell-type specific marker peaks and chromatin accessibility profiles
323 for both PBMC and solid tumor sample deconvolution. We integrated these data in the EPIC
324 deconvolution framework to accurately predict the fraction of both malignant and non-malignant cell
325 types from bulk tumor ATAC-Seq samples.
326 In cases where specific cell types are expected in a sample but are not part of our list of reference
327 profiles (*e.g.,* neuronal cells in brain tumors), custom marker peaks and reference profiles can be
328 provided to EPIC-ATAC to perform cell-type deconvolution and we provide the code to generate such
329 markers and profiles based on ATAC-Seq data from sorted cells, following the approach developed in
330 this work (Figure 1, see Code availability).
331 Solid tumors contain large and heterogeneous fractions of cancer cells for which it is
332 challenging to build reference profiles. To our knowledge this work provides the first benchmark of
333 deconvolution tools adapted to ATAC-Seq data in the context of solid tumor samples. We show that
334 the EPIC-ATAC framework, in contrast to other existing tools, allows users to accurately predict the
335 proportion of cells not included in the reference profiles (Figure 4 and Supplementary Figure 4).
336 These uncharacterized cells can include cancer cells but also other non-malignant cells. Since the
337 major cell types composing TMEs were included in our reference profiles, the proportion of
338 uncharacterized cells approximates the proportion of the cancer cells in most cases.
339 The pseudobulk approach provides unique opportunities to design benchmarks with known
340 cell-type proportions but also comes with some limitations. Indeed, pseudobulks are generated from
341 single-cell data which are noisy and whose cell-type annotation is challenging in particular for closely
342 related cell types. These limitations might lead to chromatin accessibility profiles that deviates from
343 true bulk data and errors in the true cell-type proportions. For this reason, we anticipate that the

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

351 Overall the evaluation of the EPIC-ATAC deconvolution resulted in an average absolute error
352 of 7% across cell types. This number is consistent with previous observations in RNA-Seq data
353 deconvolution (Racle et al. 2017). Considering this uncertainty, the quantification of low frequency
354 populations remains challenging (Jin and Liu 2021). While the estimated proportions of these
355 populations by EPIC-ATAC are low (*e.g.,* dendritic cells), comparing such estimations across samples
356 should be performed with care due to the uncertainty of the predictions.

357 Another limitation of cell-type deconvolution is often reached when closely related cell types
358 are considered. In the reference-based methods used in this study, this limit was reached when
359 considering T-cell subtypes in the reference profiles (Figure 5 and Supplementary figure 8). We thus
360 recommend to use the EPIC-ATAC framework using the markers and reference profiles based on the
361 major cell populations. We additionally provide the marker peaks of the T-cell subtypes which could
362 be used to build cell-type specific chromatin accessibility signatures or perform "peak set enrichment
363 analysis" similarly to gene set enrichment analysis (GSEA, (Subramanian et al. 2005)). Such
364 application could be useful for the annotation of scATAC-Seq data, which often relies on matched
365 RNA-Seq data and for which there is a lack of markers at the peak level (Jiang et al. 2023).

366 Another possible application of our marker peaks relies on their annotation (Figure 2G,
367 Supplementary Tables 4-5), which could be used to expand the list of genes and CBPs associated to
368 each cell type or subtype. For example, the neutrophils marker peaks were enriched for motifs of TFs
369 such as SPI1 (Supplementary Table 4), which was not listed in the neutrophil genes in the databases

370 used for annotation but has been reported in previous studies as involved in neutrophils
371 development (Watt et al. 2021). The annotations related to the set of major cell types and T-cell
372 subtypes are provided in Supplementary Tables 4-5 and 10-11. Finally, the annotation of marker
373 peaks highlighted pathways involved in immune responses to tumoral environments (Figure 2G).
374 Examples of these pathways are the toll-like receptor signaling pathway involved in pathogen-
375 associated and recognition of damage-associated molecular patterns in diverse cell types including B
376 and T cells (Geng et al. 2010; Javaid and Choi 2020), glucan metabolic processes which are known to
377 be related to trained immunity which can lead to anti-tumor phenotype in neutrophils (Kalafati et al.
378 2020) or the Fc-receptor signaling observed in NK cells (Sanseviero 2019; Bonnema et al. 1994).
379 These observations suggest that our marker peaks contain regulatory regions not only specific to cell
380 types but also adapted to the biological context of solid tumors.
381

382 ## Conclusion

383 In this work, we identified biologically relevant cell-type specific chromatin accessibility markers and
384 profiles for all major cancer-relevant cell types. We capitalized on these markers and profiles to
385 predict cell-type proportions from bulk PBMC and solid tumor ATAC-Seq data
386 (https://github.com/GfellerLab/EPIC-ATAC). Evaluated on diverse tissues, EPIC-ATAC shows reliable
387 predictions of immune, stromal, vascular and cancer cell proportions. With the expected increase of
388 ATAC-Seq studies in cancer, the EPIC-ATAC framework will enable researchers to deconvolve bulk
389 ATAC-Seq data from tumor samples to support the analysis of regulatory processes underlying tumor
390 development, and correlate the TME composition with clinical variables.
391

392 ## Materials and methods

393 **<u>Generation of an ATAC-Seq reference dataset of cancer relevant cell types.</u>**

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

395 We collected pure ATAC-Seq samples from 12 studies. The data include samples from (i) ten major
396 immune, stromal and vascular cell types (B (Calderon et al. 2019; Corces et al. 2016; P. Zhang et al.
397 2022), CD4+ (Corces et al. 2016; Liu et al. 2020; P. Zhang et al. 2022; Mumbach et al. 2017; Giles et al.
398 2022), CD8+ (Calderon et al. 2019; Corces et al. 2016; Liu et al. 2020; P. Zhang et al. 2022; Giles et al.
399 2022), natural killer (NK) (Calderon et al. 2019; Corces et al. 2016), dendritic (DCs) cells (Calderon et
400 al. 2019; Leylek et al. 2020; Liu et al. 2020), macrophages (Liu et al. 2020; P. Zhang et al. 2022),
401 monocytes (Calderon et al. 2019; Corces et al. 2016; Leylek et al. 2020; P. Zhang et al. 2022; Trizzino
402 et al. 2021) and neutrophils (Ram-Mohan et al. 2021; Perez et al. 2020) as well as fibroblasts (Ge et
403 al. 2021; Liu et al. 2020) and endothelial (Liu et al. 2020; Xin et al. 2020) cells (See Figure 2A), and (ii)
404 eight tissues from distinct organs (*i.e* bladder, breast, colon, liver, lung, ovary, pancreas and thyroid)
405 from the ENCODE data (The ENCODE Project Consortium et al. 2020; Rozowsky et al. 2023). The list
406 of the samples and their associated metadata (including cell types and accession number of the study
407 of origin) is provided in Supplementary Table 1. To limit batch effects, the samples were reprocessed
408 homogeneously from the raw data (fastq files) processing to the peak calling. For that purpose, raw
409 fastq files were collected from GEO using the SRA toolkit and the PEPATAC framework (Smith et al.
410 2021) was used to process the raw fastq files based on the following tool: trimmomatic for adapter
411 trimming, bowtie2 (with the PEPATAC default parameters) for reads pre-alignment on human
412 repeats and human mitochondrial reference genome, bowtie2 (with the default PEPATAC
413 parameters: *--very-sensitive -X 2000*) for alignment on the human genome (hg38), samtools
414 (PEPATAC default parameters: *-q 10*) for duplicates removal and MACS2 (Y. Zhang et al. 2008)
415 (PEPATAC default parameters: *--shift -75 --extsize 150 --nomodel --call-summits --nolambda --keep-*
416 *dup all -p 0.01*) for peak calling in each sample. After alignment, reads mapping on chromosome M
417 were excluded. TSS enrichment scores were computed for each sample and used to filter out
418 samples with low quality (criteria of exclusion: TSS score < 5) (See Supplementary Table 1 containing
419 the TSS score of each sample). 789 samples (including 564 from our ten reference cell-types) had a
420 TSS score > 5.

421

422 **Generation of a consensus set of peaks**

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

436

437 **<u>Identification of cell-type specific markers</u>**

438 **Differential accessibility analysis**

439 To identify cell-type specific markers, we split the samples collection in ten folds (created with the
440 *create_folds* function from the R package splitTools (Mayer 2023)). For each fold, we performed
441 pairwise differential accessibility analysis across the ten cell types considered in the reference
442 samples as well as the ENCODE samples from diverse organs. The differential analysis was performed
443 using limma ((Ritchie et al. 2015), version 3.56.2). Effective library sizes were computed using the
444 method of trimmed mean of M-values (TMM) from the edgeR package in R ((Robinson, McCarthy,
445 and Smyth 2010), version 3.42.4). Due to differences of library size across all samples collected, we
446 used voom from the limma package (Law et al. 2014) to transform the data and model the mean-

447 variance relationship. Finally, a linear model was fitted to the data to assess the differential
448 accessibility of each peak across each pair of cell types. To identify our marker peaks, all peaks with
449 log2 fold change higher than 0.2 were selected and ranked by their maximum adjusted *p*-value across
450 all pairwise comparisons. The top 200 features (with the lowest maximum adjusted *p*-value) were
451 considered as cell-type specific marker peaks. The marker peaks identified in at least three folds were
452 considered in the final list of marker peaks.
453
454 **Marker peaks filtering**
455 Modules of open chromatin regions accessible in all (universal modules) or in specific human tissues
456 have been identified in the study Zhang *et al.* (K. Zhang et al. 2021). These regions were used to
457 refine the set of marker peaks and exclude peaks with residual accessibility in other cell types than
458 those considered for deconvolution. More precisely, for immune, endothelial and fibroblasts specific
459 peaks, we filtered out the peaks overlapping the universal modules as well as the tissue specific
460 modules except the immune (modules 8 to 25), endothelial (modules 26 to 35) and stromal related
461 modules (modules 41 to 49 and 139-150) respectively. As a second filtering step, we retained
462 markers exhibiting the highest correlation patterns in tumor bulk samples from different cancer
463 types, *i.e.,* The Cancer Genome Atlas (TCGA) samples (Corces et al. 2018). We used the Cancer
464 Genomics Cloud (CGC) (Lau et al. 2017) to retrieve the ATAC-Seq counts for each marker peaks in
465 each TCGA sample (using *featureCounts*). For each set of cell-type specific peaks, we identified the
466 most correlated peaks using the *findCorrelation* function of the caret R package ((Kuhn 2008), version
467 6.0-94) with a correlation cutoff value corresponding to the 90th percentile of pairwise Pearson
468 correlation values.
469
470 **Evaluation of the study of origin batch effect**
471 To identify potential batch effect issues, we run principal component analysis (PCA) based on the
472 cell-type specific peaks after normalizing ATAC-Seq counts using full quantile normalization (FQ-FQ)

473 implemented in the EDASeq R package (Risso et al. 2011) to correct for depth and GC biases. These
474 data were used to visualize the data in two-dimensional space running Uniform Manifold
475 Approximation (UMAP) based on the PBMC and TME markers (Figure 2B). We also run PCA and used
476 the ten first principal components to evaluate distances between samples and compute silhouette
477 coefficients based on the cell type and study of origin classifications.
478
479 **Building the reference profiles**
480 It has been previously demonstrated in the context of RNA-Seq based deconvolution approaches
481 (Racle et al. 2017; Sturm et al. 2019) that the transcripts per million (TPM) transformation is
482 appropriate to estimate cell fractions from bulk mixtures. We thus normalized the ATAC-Seq counts
483 of the reference samples using a TPM-like transformation, *i.e.,* dividing counts by peak length,
484 correcting samples counts for depth and rescaling counts so that the counts of each sample sum to
485 10^6. We then computed for each peak the median of the TPM-like counts across all samples from
486 each cell type to build the reference profiles of the ten cell types considered in the EPIC-ATAC
487 framework (Figure 2C). In the EPIC algorithm, weights reflecting the variability of each feature of the
488 reference profile can be considered in the constrained least square optimization. We thus also
489 computed the inter-quartile range of the TPM-like counts for each feature in each cell type. Two
490 ATAC-Seq reference profiles are available in the EPIC-ATAC framework: (i) a reference profile
491 containing profiles for B cells, CD4+ T cells, CD8+ T cells, NK, monocytes, dendritic cells and
492 neutrophils to deconvolve PBMC samples, and (ii) a reference profile containing profiles for B cells,
493 CD4+ T cells, CD8+ T cells, NK, dendritic cells, macrophages, neutrophils, fibroblasts and endothelial
494 cells to deconvolve tumor samples. The reference profiles are available in the EPICATAC R package
495 and the reference profiles restricted to our cell-type specific marker peaks are available in the
496 Supplementary Tables 2 and 3.
497
498 **Assessing the reproducibility of the marker peaks signal in independent samples**

499 We evaluated the chromatin accessibility level of the marker peaks in samples that were not included
500 in the peak calling step. Firstly, we considered samples from two independent studies (Ucar et al.
501 2017; Carvalho et al. 2021) providing pure ATAC-Seq data for five immune cell types (*i.e.,* B, CD4+ T
502 cells, CD8+ T cells, Monocytes, Macrophages) (Figure 2D). To consider the other cell types, samples
503 that were excluded from the reference dataset due to a low TSS enrichment score were also
504 considered in this validation dataset (Supplementary Table 1). Secondly, we collected the data from a
505 single-cell atlas chromatin accessibility from human tissues and considered the cell types included in
506 our reference data (K. Zhang et al. 2021) (Figure 2E). We used the cell-type annotations provided in
507 the original study (GEO accession number: GSE184462). The Signac R package ((Stuart et al. 2021),
508 1.9.0) was used to extract fragments counts for each cell and each marker peak and the ATAC-Seq
509 signal of each marker peak was averaged across all cells of each cell type.
510
511 **Annotation of the marker peaks**
512 The cell-type specific markers were annotated using ChIPseeker R package ((Yu, Wang, and He 2015),
513 version 1.34.1) and the annotation from *TxDb.Hsapiens.UCSC.hg38.knownGene* in R to identify the
514 regions in which the marker peaks are (*i.e.,* promoter, intronic regions, etc.) and ChipEnrich to
515 associate each peak to the nearest gene TSS (Welch et al. 2014). The nearest genes identified were
516 then compared to cell-type marker genes listed in the PanglaoDB (Franzén, Gan, and Björkegren
517 2019) and CellMarker databases (Hu et al. 2023). PanglaoDB provides an online interface to explore a
518 large collection on single-cell RNA-Seq data as well as a community-curated list of cell-type marker
519 genes. CellMarker is a database providing a large set of curated cell-type markers for more than 400
520 cell types in human tissues retrieved from a large collection of single-cell studies and flow cytometry,
521 immunostaining or experimental studies. ChipEnrich was also used to perform gene set enrichment
522 and identify for each set of cell-type specific peaks potential biological pathways regulated by the
523 marker peaks. The enrichment analysis was performed using the *chipenrich* function (*genesets =*
524 *"GOBP" , locusdef = "nearest_tss"*) from the chipenrich R package (v2.22.0).

525 Chromatin accessibility peaks can also be annotated for chromatin binding proteins (CBPs) such as
526 transcription factors (TFs), whose potential binding in the peak region is reported in databases. In our
527 study we chose the JASPAR2022 (Castro-Mondragon et al. 2022) database and the ReMap database
528 (Hammal et al. 2022).
529 Using the JASPAR2022 database, we assessed, for each cell type, whether the cell-type specific
530 marker peaks were enriched in specific TFs motifs using two TFs enrichment analysis frameworks:
531 Signac (Stuart et al. 2021) and MonaLisa (Machlab et al. 2022). For the MonaLisa analysis, the cell-
532 types specific markers peaks were categorized in bins of sequences, one bin per cell type (use of the
533 *calcBinnedMotifEnrR* function). To test for an enrichment of motifs, the sequences of each bin were
534 compared to a set of background peaks with similar average size and GC composition obtained by
535 randomly sampling regions in all the peaks identified from the reference dataset. The enrichment
536 test was based on a binomial test. For the Signac analysis, we used the *FindMotif* function to identify
537 over-represented TF motifs in each set of cell-type specific marker peaks (query). This function used a
538 hypergeometric test to compare the number of query peaks containing the motif with the total
539 number of peaks containing the motif in the background regions (matched to have similar GC
540 content, region length and dinucleotide frequencies as the query regions), corresponding in our case
541 to the peaks called in the reference dataset.
542 The ReMap database associates chromatin binding proteins (CBPs), including TFs, transcriptional
543 coactivators and chromatin-remodeling factors, to their DNA binding regions based on DNA-binding
544 experiments such as chromatin immunoprecipitation followed by sequencing (ChIP-seq). For each
545 association of a CBP to its binding region, the cell type in which the binding has been observed is
546 reported in the ReMap database (biotype). We used the ReMapEnrich R package (version 0.99) to
547 test if the cell-type specific marker peaks are significantly enriched in CBPs-binding regions listed in
548 the Remap 2022 catalog. We considered the non-redundant peaks catalog from Remap 2022,
549 containing non-redundant binding regions for each CBP in each biotype. Similarly to the previously
550 mentioned enrichment methods, we chose the consensus peaks called in the reference samples as

551 universe for the enrichment test. Note that, for each cell type, an enrichment was retained only if the
552 biotype in which the CBP-regions were identified matched the correct cell-type.
553
554 **Running EPIC-ATAC on bulk ATAC-Seq data**
555 The samples used to generate the reference profiles were aligned using the hg38 reference genome.
556 To assure the compatibility of any input bulk ATAC-Seq dataset with the EPIC-ATAC marker peaks and
557 reference profiles, we provide an option to lift over hg19 datasets to hg38 (use of the liftOver R
558 package). Subsequently, the features of the input bulk matrix are matched to our reference profiles
559 features. To match both sets of features, we determine for each peak of the input bulk matrix the
560 distance to the nearest peak in the reference profiles peaks. Overlapping regions are retained and
561 the feature IDs are matched to their associated nearest peaks. If multiple features are matched to
562 the same reference peak, the counts are summed. In RNA-Seq based deconvolution, EPIC uses an
563 estimation of the amount of mRNA in each reference cell type to derive cell proportions. For the
564 ATAC-Seq based deconvolution these values were set to 1 to give similar weights to all cell-types
565 quantifications.
566
567 **Datasets used for the evaluation of ATAC-Seq deconvolution**
568 **PBMCs ATAC-Seq data from healthy donors**
569 ***Peripheral blood mononuclear cell (PBMC) isolation***
570 Venous blood from five healthy donors was collected at the local blood transfusion center of Geneva
571 in Switzerland, under the approval of the Geneva University Hospital's Institute Review Board, upon
572 written informed consent and in accordance with the Declaration of Helsinki. PBMCs were freshly
573 isolated by Lymphoprep (Promega) centrifugation (1800 rpm, 20 minutes, without break, room
574 temperature). Red blood cell lysis was performed using red blood lysis buffer (Qiagen) and platelets
575 were removed by centrifugation (1000 rpm, 10 minutes without break, room temperature). Cells
576 were counted and immediately used.

577

578 ***Flow cytometry***

579 Immune cell populations were identified using multiparameter flow cytometry and the following
580 antibodies: FITC anti-human CD45RA (HI100, Biolegend), PerCP-Cyanine5.5 anti-human CD19 (H1B19,
581 Biolegend), PE anti-human CD3 (SK7, Biolegend), PE-Dazzle anti-human CD14 (M0P9, BD
582 Biosciences), PE-Cyanine7 anti-human CD56 (HCD56, Biolegend), APC anti-human CD4 (RPA-T4,
583 Biolgend), APC-Cyanine7 anti-human CCR7 (G043H7, Biolegend), Brilliant Violet 421 anti-human CD8
584 (RPA-T8, Biolegend), Brilliant Violet 510 anti-human CD25 (BC96, Biolegend), Brilliant Violet 711 anti-
585 human CD16 (3G8, Biolegend), Brilliant Violet 786 anti-human CD127 (A019D5, Biolegend), Ultra-
586 Brilliant Violet anti-human CD45 (HI30, BD Biosciences), FITC anti-human Celc9a (8F9, Miltenyi) , PE
587 anti-human XCR1 (S15046E, Biolegend), PE-Dazzle anti-human BDCA-2 (201A, Biolegend), APC anti-
588 human BDCA-3 (AD5-14H12, Miltenyi), Brilliant Violet 421 anti-human CD3 (UCHT1, Biolegend),
589 Brilliant Violet 421 anti-human CD14 (M5E2, BD Pharmingen), Brilliant Violet 421 anti-human CD19
590 (SJ25C1, Biolegend), Brilliant Violet 510 anti-human BDCA-1 (L161, Biolegend), Brilliant Violet 650
591 anti-human CD11c (3.9, Biolegend), Brilliant Violet 711 anti-human CD11c (N418, Biolegend) and
592 Brilliant Violet 711 anti-human HLA-DR (L243, Biolegend). Dead cells were excluded using the Zombie
593 UV™ Fixable Viability Kit (Biolegend). Intracellular staining was performed after fixation and
594 permeabilization of the cells with the FoxP3 Transcription Factor Staining Buffer Set (00-5523-00,
595 Invitrogen) using Alexa 700 anti-human FoxP3 antibody (259D/C7, BD Biosciences). Data were
596 acquired on LSRFortessa flow cytometer and analysed using FlowJo software (v10.7.1).

597

598 ***Cell preparation for ATAC-Sequencing***

599 50000 CD45+ cells were sorted from total PBMCs using anti-human Ultra-Brilliant Violet (BUV395)
600 CD45 (HI30, BD Biosciences) with a FACSAria II (Becton Dickinson) and were collected in PBS with
601 10% Foetal Bovine Serum (FBS). Cell pellets were resuspended in cold lysis buffer (10mM Tris-Cl pH
602 7.4, 10mM NaCl, 3mM MgCl2, 0,1% NP40 and water) and immediately centrifuged at 600g for 30min

603 at 4°C. Transposition reaction was performed using the Illumina Tagment DNA Enzyme and Buffer kit
604 (20034210, Illumina) and transposed DNA was eluted using the MinElute PCR Purification Kit
605 (Qiagen). Libraries were generated by PCR amplification using indexing primers and NEBNext High-
606 Fidelity Master Mix (New England BioLabs) and were purified using AMPure XP beads (A63880,
607 Beckman Coulter). Libraries were quantified by a fluorometric method (QubIT, Life Technologies) and
608 their quality assessed on a Fragment Analyzer (Agilent Technologies). Sequencing was performed as a
609 paired end 50 cycles run on an Illumina NovaSeq 6000 (v1.5 reagents) at the Genomic Technologies
610 Facility (GTF) in Lausanne, Switzerland. Raw sequencing data were demultiplexed using the
611 bcl2fastq2 Conversion Software (version 2.20, Illumina).
612
613 ***Data processing***
614 The same steps as for the processing of the reference ATAC-Seq samples were followed. (See Pre-
615 processing of the ATAC-Seq datasets).
616
617 **ATAC-Seq pseudobulk data from PBMCs and cancer samples**
618 To evaluate the accuracy of our ATAC-Seq deconvolution framework, we generated pseudo-bulk
619 datasets from 5 single-cell datasets:
620 - **PBMC pseudobulk dataset:** combination of three single-cell datasets for PBMCs.
621 - Dataset 1 corresponds to a scATAC-Seq dataset obtained from Satpathy et al.
622 (Satpathy et al. 2019) (GEO accession number: GSE129785). This dataset contains
623 FACS-sorted populations of PBMCs. Since the cells of some cell types came from a
624 unique donor, all the cells of this dataset were aggregated to form one pseudobulk.
625 Ground truth cell fractions were obtained by dividing the number of cells in each cell
626 type by the total number of cells.
627 - Dataset 2 (included in the PBMC pseudobulk dataset) was retrieved from Granja *et*
628 *al.* (Granja et al. 2019) (GEO accession number GSE139369). B cells, monocytes,

629 dendritic, CD8+, CD4+ T, NK cells, neutrophils from healthy donors were considered.
630 The neutrophil cells came from a single donor. As for dataset 1, we thus aggregated
631 all the cells to generate one pseudobulk. Ground truth cell fractions were obtained
632 by dividing the number of cells in each cell type by the total number of cells.

633 - Dataset 3 (included in the PBMC pseudobulk dataset) corresponds to the 10X
634 multiome dataset of PBMC cells (10x Genomics 2021). Since these data come from
635 one donor, one pseudobulk sample was generated for this dataset. The pseudobulk
636 was generated by averaging the ATAC-Seq signal from all cells from the following cell
637 types: B cells, CD4+ T cells , CD8+ T cells, NK cells, Dendritic cells and monocytes.

638 - **Basal cell carcinoma dataset:** obtained from the study of Satpathy *et al.* (Satpathy et al.
639 2019). This dataset is a scATAC-Seq dataset composed of 13 basal cell carcinoma samples
640 composed of immune (B cells, plasma cells, CD4+ T cells, CD8+ T cells, NK cells, myeloid cells),
641 stromal (endothelial and fibroblasts) and cancer cells. Plasma cells and cancer cells were both
642 considered as uncharacterized cells (*i.e.,* cell types not included in the reference profiles).
643 Cell annotations were retrieved from the original study.

644 - **Gynecological cancer dataset:** obtained from the study of Regner *et al.* (Regner et al. 2021)
645 (GEO accession number GSE173682). In this study, the authors performed scATAC-Seq on 11
646 gynecological cancer samples from two tumor sites (*i.e* endometrium and ovary) and
647 composed of immune (B cells, NK and T cells grouped under the same cell-type annotation,
648 macrophages, mast cells), stromal (fibroblast, endothelial, smooth muscle) and cancer cells.
649 Mast cells, smooth muscle and cancer cells were considered as uncharacterized cells. Cell
650 annotations were retrieved from the original study.

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

654 For each dataset, raw fragments files were downloaded from the respective GEO accession numbers
655 and data were preprocessed using ArchR ((Granja et al. 2021), ArchR R package 1.0.2). Cells with TSS
656 score below four were removed. Doublets removal was performed using the *doubletsRemoval*
657 function from ArchR. To match as much as possible real bulk ATAC-seq data processing, peak calling
658 was not performed on each cell type or cell cluster as usually done in scATAC-Seq studies but using
659 all cells for each dataset from the PBMC pseudobulk data or grouping cells by sample for the Basal
660 cell carcinoma and Gynecological cancer datasets. Peak calling was performed using MACS2 within
661 the ArchR framework. Fragments counts were extracted using ArchR for each peak called to generate
662 single-cell peak counts matrices. These matrices were normalized using a TPM-like transformation,
663 *i.e.,* dividing counts by peak length and correcting samples counts for depth. Finally, for each peak,
664 the average of the normalized counts was computed across all the cells for each dataset from the
665 PBMC pseudobulk data and across all the cells of each sample for the Basal cell carcinoma and
666 Gynecological cancer datasets. Averaged data were then rescaled so that the sum of counts of each
667 sample sum to 10^6.
668
669 **Bulk ATAC-Seq data from a breast cancer cohort**
670 Bulk ATAC-Seq samples from a breast cancer cohort was obtained from Kumegawa *et al.* (Kumegawa
671 et al. 2023). These data include 42 breast cancer samples which can be classified based on two
672 features: (i) the breast cancer subtype ER+/HER2- or triple negative, and (ii) the molecular
673 classification provided by the original study (CA-A, CA-B and CA-C). The ATAC-Seq raw counts and the
674 samples metadata were retrieved from figshare (Kumegawa 2023). As for the previously mentioned
675 datasets, raw counts were normalized using the TPM-like transformation prior to bulk deconvolution.
676
677 **<u>Benchmarking of the EPIC-ATAC framework against other existing deconvolution tools</u>**
678 The performances of the EPIC-ATAC framework were benchmarked against the following
679 deconvolution tools:

680 • quanTIseq (Finotello et al. 2019) is a deconvolution tool using constrained least square
681 regression to deconvolve RNA-Seq bulk samples. No reference profiles are available in this
682 framework to perform ATAC-Seq deconvolution and quanTIseq does not provide the option
683 to automatically build reference profiles from pure bulk samples. quanTIseq was thus run
684 using the reference profiles derived in this work for the EPIC-ATAC framework and the
685 *quanTIseq* function from the quantiseqr R package (parameters: scaling set to 1 for all cell
686 types and method set to "lsei").
687 • DeconPeaker (H. Li et al. 2020) relies on SIMPLS, a variant of partial least square regression
688 to perform bulk RNA-Seq and bulk ATAC-Seq deconvolution. ATAC-Seq reference profiles are
689 available in this deconvolution framework however not all cell types considered in the EPIC-
690 ATAC framework are included in the DeconPeaker reference profiles. This tool was thus run
691 using different reference profiles: (i) the reference profiles derived in this work for the EPIC-
692 ATAC framework (corresponds to “DeconPeaker” or “DeconPeaker_ourmarkers” in our
693 analyses), and (ii) reference profiles automatically generated by DeconPeaker from the
694 sorted reference samples collected in this work (corresponds to “DeconPeaker_cust.” in our
695 analyses). The results of DeconPeaker obtained using its original markers and profiles are
696 also provided for the cell types in common with the cell types considered in this work in
697 Supplementary Figures 3 and 5. Deconvolution was run using the deconvolution module
698 deconPeaker (using findctsps with the following parameter: --lib-strategy=ATAC-Seq).
699 DeconPeaker outputs cell-type proportions relative to the total amount of cells from the
700 reference cell types.
701 • CIBERSORTx (Newman et al. 2019) is a deconvolution algorithm based on linear support
702 vector regression. CIBERSORTx does not provide ATAC-Seq reference profiles, however it is
703 possible to automatically generate new profiles from a set of pure bulk samples. This tool
704 was thus run using different reference profiles: i) the reference profiles derived in this work
705 for the EPIC-ATAC framework (corresponds to “CIBERSORTx” or “CIBERSORTx_ourMarkers”

706 in our analyses), and ii) reference profiles automatically generated by CIBERSORTx from the
707 sorted reference samples collected in this work (corresponds to "CIBERSORTx_cust." in our
708 analyses). To run CIBERSORTx, we used the docker container provided by the authors of
709 CIBERSORTx on their website. The algorithm was run using the default options (i.e --absolute
710 FALSE, --rmbatchBmode FALSE and –rmbatchSmode FALSE), which results in cell-type
711 proportions relative to the total amount of cells from the reference cell types.

712 • ABIS (Monaco et al. 2019) uses robust linear modeling to estimate cell-type proportions in
713 bulk RNA-Seq samples. No ATAC-Seq reference profiles are available in the deconvolution
714 framework. ABIS was run using the EPIC-ATAC reference profiles by using the *rlm* function
715 from the MASS R package (as performed in the deconvolute_abis function from the
716 immunedeconv R package (Sturm et al. 2019) was used to quantify each cell type from the
717 reference profiles. The cell-types quantifications returned by this approach are in arbitrary
718 units. To compare the estimations and the true cell proportions, we scaled the estimations of
719 each sample between 0 and 1 to obtained relative proportions.

720 • MCPcounter (Becht et al. 2016): MCPcounter returns scores instead of cell type proportions.
721 The scores were obtained using the *appendSignatures* function from the MCPcounter R
722 package by providing the list of marker peaks specific to each cell type. The cell-type scores
723 are not comparable between cell type, MCPcounter was thus included only in the evaluation
724 of the performances in each cell type separately.

725

726 For all the tools, TPM-like data were used as input bulk samples for the deconvolution.
727 Since CIBERSORTx, ABIS and DeconPeaker do not predict proportions of uncharacterized cells, we
728 performed two benchmarking analyses: (i) including all cell types and (ii) excluding the cell types that
729 are absent from the reference profiles (uncharacterized cells) and rescaling the estimated and true
730 proportions of the immune cells, endothelial cells and fibroblasts so that their sum equals 1.

731

732 **Comparing deconvolution based on RNA-Seq, gene activity or peaks features.**

733 100 pseudobulks were generated from the 10X PBMC multiome dataset (10x Genomics 2021) based
734 on 3000 cells for each pseudobulk. Cell fractions were defined using the *rdirichlet* function from the
735 *gtools* R package. Three sets of features were extracted from the data, *i.e.,* gene expression features
736 extracted from the RNA-Seq layer, ATAC-Seq peaks and gene activity derived from the ATAC-Seq
737 layer. The same cells sampling was considered for each modality.

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

748

749 **Code availability**

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

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

756

757 **Data availability**

758 The newly generated ATAC-Seq data have been deposited on Zenodo (doi:
759 10.5281/zenodo.8431792). The other data related to this work are available in the supplementary
760 tables and on the Zenodo deposit (doi: 10.5281/zenodo.8431792).
761
762 **Competing interests:**
763 The authors declare that they have no competing interests.
764
765 **Acknowledgement:**
766 We thank the Lausanne Genomic Technologies Facility, University of Lausanne, Switzerland
767 (https://www.unil.ch/gtf/en/home.html) for the sequencing of the PBMC samples as well as Yan Liu,
768 Dana Moreno and Matei Teleman for testing the EPICATAC R package. Some of the illustrations were
769 created with BioRender.com.
770
771 **Authors contributions:**
772 Conceptualization: AAGG, JR, DG; Data curation: AAGG; Software: AAGG, JR, DG; Experiments: MF,
773 CJ; Visualization: AAGG, JR, DG; Methodology: AAGG, JR, DG; Writing—original draft: AAGG, DG;
774 Writing—review and editing: all authors.
775
776 **Figure legends:**
777 ***Figure 1: Graphical description of the identification of cell-type specific marker peaks and reference***
778 ***ATAC-Seq profiles included in the EPIC-ATAC framework.*** *1) 564 pure ATAC-Seq data of sorted cells*
779 *were collected to build reference profiles for cancer-relevant cell populations. 2) Cell-type specific*
780 *marker peaks were identified using differential accessibility analysis. 3) Markers with previously*
781 *observed chromatin accessibility in human healthy tissues were then excluded. 4) For tumor bulk*
782 *deconvolution, the set of remaining marker peaks was refined by selecting markers with correlated*
783 *behavior in tumor bulk samples. 5) The cell-type specific marker peaks and reference profiles were*

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

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

Figure 3: EPIC-ATAC accurately estimates immune cell fractions in PBMC ATAC-Seq samples. A) *Schematic description of the experiment designed to validate the ATAC-Seq deconvolution on PBMC samples.* ***B)*** *Comparison between cell-type proportions predicted by EPIC-ATAC and the true proportions in the PBMC bulk dataset. Symbols correspond to donors.* ***C)*** *Comparison between the proportions of cell-types predicted by EPIC-ATAC and the true proportions in the PBMC pseudobulk dataset. Symbols correspond to pseudobulks.* ***D)*** *Pearson correlation (left) and RMSE (right) values obtained by each deconvolution tool on the PBMC bulk dataset. The EPIC-ATAC results are highlighted*

810 *in red.* ***E)*** *Pearson correlation (left) and RMSE (right) values obtained by each deconvolution tool on*
811 *the PBMC pseudobulk dataset. Parts of this figure (panel 1) were created with BioRender.com.*
812
813 ***Figure 4: EPIC-ATAC accurately predicts fractions of cancer and non-malignant cells in tumor***
814 ***samples. A)*** *Comparison between cell-type proportions estimated by EPIC-ATAC and true proportions*
815 *for the basal cell carcinoma (top) and gynecological (bottom) pseudobulk datasets. Symbols*
816 *correspond to pseudobulks.* ***B)*** *Pearson's correlation and RMSE values obtained for the deconvolution*
817 *tools included in the benchmark. EPIC-ATAC is highlighted in red.* ***C)*** *Same analyses as in panels B,*
818 *with the uncharacterized cell population excluded for the evaluation of the predictions accuracy. The*
819 *predicted and true proportions of the immune, stromal and vascular cell types were rescaled to sum*
820 *to 1.*
821
822 ***Figure 5: T cell subtypes quantification reveals the ATAC-Seq deconvolution limits for closely***
823 ***related cell types. A)*** *Comparison of the proportions estimated by EPIC-ATAC and the true proportions*
824 *for PBMC samples (PBMC experiment and PBMC pseudobulk samples combined) (top) and the basal*
825 *cell carcinoma pseudobulks (bottom). Predictions of the proportions of CD4+ and CD8+ T-cells were*
826 *obtained using the reference profiles based on the major cell types and subtype predictions using the*
827 *reference profiles including the T-cell subtypes.* ***B)*** *Pearson's correlation values obtained by EPIC-ATAC*
828 *in each cell type.*
829
830 ***Figure 6: EPIC-ATAC accurately infers the immune contexture in a bulk ATAC-Seq breast cancer***
831 ***cohort. A)*** *Proportions of different cell types predicted by EPIC-ATAC in the samples stratified based*
832 *on two breast cancer subtypes.* ***B)*** *Proportions of different cell types predicted by EPIC-ATAC in the*
833 *samples stratified based on three ER+/HER2- subgroups. Wilcoxon test p-values are represented at*
834 *the top of the boxplots.*
835

836 ***Figure 7: EPIC-ATAC performs similarly to EPIC RNA-seq based deconvolution and better than gene***
837 ***activity based deconvolution.*** *Pearson's correlation (left) and RMSE (right) values comparing the*
838 *proportions predicted by the ATAC-Seq deconvolution, the RNA-Seq deconvolution and the GA-based*
839 *RNA deconvolution and true cell-type proportions in the 100 pseudobulks simulated form the 10x*
840 *multiome PBMC dataset* (10x Genomics 2021)*. Dots correspond to outlier pseudobulks.*
841
842 **Supplementary Figures:** Additional file named Supplementary_figures.pdf
843 **Supplementary Tables:**
844 **Sup. Table 1:** Metadata of the ATAC-Seq samples used in the study
845 **Sup. Table 2:** Averaged chromatin accessibility of the PBMC marker peaks in each cell-type.
846 **Sup. Table 3:** Averaged chromatin accessibility of the TME marker peaks in each cell-type.
847 **Sup. Table 4:** Annotations of the cell-type specific PBMC marker peaks
848 **Sup. Table 5:** Annotations of the cell-type specific TME marker peaks
849 **Sup. Table 6:** GO pathways enriched in each set of cell-type specific PBMC marker peaks
850 **Sup. Table 7:** GO pathways enriched in each set of cell-type specific TME marker peaks
851 **Sup. Table 8:** Averaged chromatin accessibility of the PBMC marker peaks in each cell-type (T cells
852 subtypes included).
853 **Sup. Table 9:** Averaged chromatin accessibility of the TME marker peaks in each cell-type (T cells
854 subtypes included).
855 **Sup. Table 10:** Annotations of the cell-type specific PBMC marker peaks (T cells subtypes included).
856 **Sup. Table 11:** Annotations of the cell-type specific TME marker peaks (T cells subtypes included).
857 **Sup. Table 12:** GO pathways enriched in each set of cell-type specific PBMC marker peaks (T cell
858 subtypes).
859 **Sup. Table 13:** GO pathways enriched in each set of cell-type specific TME marker peaks (T cell
860 subtypes).
861

862 **References:**

863

864 10x Genomics. 2021. "PBMC from a Healthy Donor - Granulocytes Removed Through Cell Sorting
865 (10k)." 2021. https://www.10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donor-
866 granulocytes-removed-through-cell-sorting-10-k-1-standard-2-0-0.

867 Arneson, Douglas, Xia Yang, and Kai Wang. 2020. "MethylResolver—a Method for Deconvoluting Bulk
868 DNA Methylation Profiles into Known and Unknown Cell Contents." *Communications Biology* 3.
869 https://doi.org/10.1038/s42003-020-01146-2.

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

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

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

880 Bonnema, Joy D., Larry M. Karnitz, Renee A. Schoon, Robert T. Abraham, and Paul J. Leibson. 1994.
881 "Fc Receptor Stimulation of Phosphatidylinositol 3-Kinase in Natural Killer Cells Is Associated
882 with Protein Kinase C-Independent Granule Release and Cell-Mediated Cytotoxicity." *Journal of*
883 *Experimental Medicine* 180 (4): 1427–35. https://doi.org/10.1084/JEM.180.4.1427.

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

888 Burdziak, Cassandra, Elham Azizi, Sandhya Prabhakaran, and Dana Pe'er. 2019. "A Nonparametric
889 Multi-View Model for Estimating Cell Type-Specific Gene Regulatory Networks." *ArXiv*.
890 Calderon, Diego, Michelle L.T. Nguyen, Anja Mezger, Arwa Kathiria, Fabian Müller, Vinh Nguyen,
891 Ninnia Lescano, et al. 2019. "Landscape of Stimulation-Responsive Chromatin across Diverse
892 Human Immune Cells." *Nature Genetics* 51 (10): 1494–1505. https://doi.org/10.1038/s41588-
893 019-0505-9.
894 Carvalho, Klebea, Elisabeth Rebboah, Camden Jansen, Katherine Williams, Andrew Dowey, Cassandra
895 McGill, and Ali Mortazavi. 2021. "Uncovering the Gene Regulatory Networks Underlying
896 Macrophage Polarization Through Comparative Analysis of Bulk and Single-Cell Data." *BioRxiv*,
897 January. https://doi.org/10.1101/2021.01.20.427499.
898 Castro-Mondragon, Jaime A., Rafael Riudavets-Puig, Ieva Rauluseviciute, Roza Berhanu Lemma, Laura
899 Turchi, Romain Blanc-Mathieu, Jeremy Lucas, et al. 2022. "JASPAR 2022: The 9th Release of the
900 Open-Access Database of Transcription Factor Binding Profiles." *Nucleic Acids Research* 50 (D1):
901 D165–73. https://doi.org/10.1093/NAR/GKAB1113.
902 Chakravarthy, Ankur, Andrew Furness, Kroopa Joshi, Ehsan Ghorani, Kirsty Ford, Matthew J. Ward,
903 Emma V. King, et al. 2018. "Pan-Cancer Deconvolution of Tumour Composition Using DNA
904 Methylation." *Nature Communications* 9 (August). https://doi.org/10.1038/s41467-018-05570-
905 1.
906 Clarke, Jennifer, Pearl Seol, and Bertrand Clarke. 2010. "Statistical Expression Deconvolution from
907 Mixed Tissue Samples." *Bioinformatics* 26 (8): 1043–49.
908 https://doi.org/10.1093/BIOINFORMATICS/BTQ097.
909 Corces, M. Ryan, Jason D. Buenrostro, Beijing Wu, Peyton G. Greenside, Steven M. Chan, Julie L.
910 Koenig, Michael P. Snyder, et al. 2016. "Lineage-Specific and Single-Cell Chromatin Accessibility
911 Charts Human Hematopoiesis and Leukemia Evolution." *Nature Genetics* 48 (August): 1193–
912 1203. https://doi.org/10.1038/ng.3646.
913 Corces, M. Ryan, Jeffrey M. Granja, Shadi Shams, Bryan H. Louie, Jose A. Seoane, Wanding Zhou,

Tiago C. Silva, et al. 2018. “The Chromatin Accessibility Landscape of Primary Human Cancers.” *Science* 362 (6413). https://doi.org/10.1126/science.aav1898.

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

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

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

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

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

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

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

940 Ge, Xiangyu, Mojca Frank-Bertoncelj, Kerstin Klein, Amanda McGovern, Tadeja Kuret, Miranda
941 Houtman, Blaž Burja, et al. 2021. "Functional Genomics Atlas of Synovial Fibroblasts Defining
942 Rheumatoid Arthritis Heritability." *Genome Biology* 22 (1): 247.
943 https://doi.org/10.1186/S13059-021-02460-6/FIGURES/7.
944 Geng, Degui, Liqin Zheng, Ratika Srivastava, Nicole Asprodites, Cruz Velasco-Gonzalez, and Eduardo
945 Davila. 2010. "When Toll-like Receptor and T-Cell Receptor Signals Collide: A Mechanism for
946 Enhanced CD8 T-Cell Effector Function." *Blood* 116 (18): 3494–3504.
947 https://doi.org/10.1182/BLOOD-2010-02-268169.
948 Giles, Josephine R., Sasikanth Manne, Elizabeth Freilich, Derek A. Oldridge, Amy E. Baxter, Sangeeth
949 George, Zeyu Chen, et al. 2022. "Human Epigenetic and Transcriptional T Cell Differentiation
950 Atlas for Identifying Functional T Cell-Specific Enhancers." *Immunity* 55 (3): 557-574.e7.
951 https://doi.org/10.1016/J.IMMUNI.2022.02.004.
952 Gong, Ting, and Joseph D. Szustakowski. 2013. "DeconRNASeq: A Statistical Framework for
953 Deconvolution of Heterogeneous Tissue Samples Based on MRNA-Seq Data." *Bioinformatics* 29
954 (8): 1083–85. https://doi.org/10.1093/BIOINFORMATICS/BTT090.
955 Gosink, Mark M., Howard T. Petrie, and Nicholas F. Tsinoremas. 2007. "Electronically Subtracting
956 Expression Patterns from a Mixed Cell Population." *Bioinformatics* 23 (24): 3328–34.
957 https://doi.org/10.1093/BIOINFORMATICS/BTM508.
958 Grandi, Fiorella C., Hailey Modi, Lucas Kampman, and M. Ryan Corces. 2022. "Chromatin Accessibility
959 Profiling by ATAC-Seq." *Nature Protocols*, April, 1518–52. https://doi.org/10.1038/s41596-022-
960 00692-9.
961 Granja, Jeffrey M., M. Ryan Corces, Sarah E. Pierce, S. Tansu Bagdatli, Hani Choudhry, Howard Y.
962 Chang, and William J. Greenleaf. 2021. "ArchR Is a Scalable Software Package for Integrative
963 Single-Cell Chromatin Accessibility Analysis." *Nature Genetics* 53 (February): 403–11.
964 https://doi.org/10.1038/s41588-021-00790-6.
965 Granja, Jeffrey M., Sandy Klemm, Lisa M. McGinnis, Arwa S. Kathiria, Anja Mezger, M. Ryan Corces,

Benjamin Parks, et al. 2019. “Single-Cell Multiomic Analysis Identifies Regulatory Programs in Mixed-Phenotype Acute Leukemia.” *Nature Biotechnology* 37 (12): 1458–65. https://doi.org/10.1038/s41587-019-0332-7.

Hammal, Fayrouz, Pierre De Langen, Aurelie Bergon, Fabrice Lopez, and Benoit Ballester. 2022. “ReMap 2022: A Database of Human, Mouse, Drosophila and Arabidopsis Regulatory Regions from an Integrative Analysis of DNA-Binding Sequencing Experiments.” *Nucleic Acids Research* 50 (D1): D316–25. https://doi.org/10.1093/NAR/GKAB996.

Hu, Congxue, Tengyue Li, Yingqi Xu, Xinxin Zhang, Feng Li, Jing Bai, Jing Chen, et al. 2023. “CellMarker 2.0: An Updated Database of Manually Curated Cell Markers in Human/Mouse and Web Tools Based on ScRNA-Seq Data.” *Nucleic Acids Research* 51 (D1): D870–76. https://doi.org/10.1093/NAR/GKAC947.

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

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

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

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

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

992 Klemm, Sandy L., Zohar Shipony, and William J. Greenleaf. 2019. “Chromatin Accessibility and the
993 Regulatory Epigenome.” *Nature Reviews Genetics* 20 (4): 207–20.
994 https://doi.org/10.1038/s41576-018-0089-8.
995 Kuhn, Max. 2008. “Building Predictive Models in R Using the Caret Package.” *Journal of Statistical*
996 *Software* 28 (5): 1–26. https://doi.org/10.18637/JSS.V028.I05.
997 Kumegawa, Kohei. 2023. “ATAC-Seq Data of 42 BC Samples as SummarizedExperiment Object with
998 Count Matrix, Normalized Count Matrix, Peak Info, and Clinical Info.” 2023.
999 https://doi.org/10.6084/m9.figshare.21992609.v1.
1000 Kumegawa, Kohei, Sumito Saeki, Yoko Takahashi, Liying Yang, Tomo Osako, Tomoyoshi Nakadai,
1001 Sayuri Amino, et al. 2023. “Chromatin Profile-Based Identification of a Novel ER-Positive Breast
1002 Cancer Subgroup with Reduced ER-Responsive Element Accessibility.” *British Journal of Cancer*
1003 128 (7): 1208–22. https://doi.org/10.1038/s41416-023-02178-1.
1004 Lareau, Caleb A., Fabiana M. Duarte, Jennifer G. Chew, Vinay K. Kartha, Zach D. Burkett, Andrew S.
1005 Kohlway, Dmitry Pokholok, et al. 2019. “Droplet-Based Combinatorial Indexing for Massive-
1006 Scale Single-Cell Chromatin Accessibility.” *Nature Biotechnology* 37 (8): 916–24.
1007 https://doi.org/10.1038/s41587-019-0147-6.
1008 Lau, Jessica W., Erik Lehnert, Anurag Sethi, Raunaq Malhotra, Gaurav Kaushik, Zeynep Onder, Nick
1009 Groves-Kirkby, et al. 2017. “The Cancer Genomics Cloud: Collaborative, Reproducible, and
1010 Democratized - A New Paradigm in Large-Scale Computational Research.” *Cancer Research* 77
1011 (21): e3–6. https://doi.org/10.1158/0008-5472.CAN-17-0387.
1012 Law, Charity W., Yunshun Chen, Wei Shi, and Gordon K. Smyth. 2014. “Voom: Precision Weights
1013 Unlock Linear Model Analysis Tools for RNA-Seq Read Counts.” *Genome Biology* 15 (February):
1014 R29. https://doi.org/10.1186/GB-2014-15-2-R29/FIGURES/11.
1015 Leylek, Rebecca, Marcela Alcántara-Hernández, Jeffrey M. Granja, Michael Chavez, Kimberly Perez,
1016 Oscar R. Diaz, Rui Li, Ansuman T. Satpathy, Howard Y. Chang, and Juliana Idoyaga. 2020.
1017 “Chromatin Landscape Underpinning Human Dendritic Cell Heterogeneity.” *Cell Reports* 32 (12):

1018 108180. https://doi.org/10.1016/J.CELREP.2020.108180.

1019 Li, Huamei, Amit Sharma, Kun Luo, Zhaohui S. Qin, Xiao Sun, and Hongde Liu. 2020. "DeconPeaker, a
1020 Deconvolution Model to Identify Cell Types Based on Chromatin Accessibility in ATAC-Seq Data
1021 of Mixture Samples." *Frontiers in Genetics* 11 (June).
1022 https://doi.org/10.3389/fgene.2020.00392.

1023 Li, Taiwen, Jingxin Fu, Zexian Zeng, David Cohen, Jing Li, Qianming Chen, Bo Li, and X. Shirley Liu.
1024 2020. "TIMER2.0 for Analysis of Tumor-Infiltrating Immune Cells." *Nucleic Acids Research* 48
1025 (May): W509–14. https://doi.org/10.1093/nar/gkaa407.

1026 Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. "FeatureCounts: An Efficient General Purpose
1027 Program for Assigning Sequence Reads to Genomic Features." *Bioinformatics* 30 (7): 923–30.
1028 https://doi.org/10.1093/BIOINFORMATICS/BTT656.

1029 Liu, Qian, Lisa Zaba, Ansuman T. Satpathy, Michelle Longmire, Wen Zhang, Kun Li, Jeffrey Granja, et
1030 al. 2020. "Chromatin Accessibility Landscapes of Skin Cells in Systemic Sclerosis Nominate
1031 Dendritic Cells in Disease Pathogenesis." *Nature Communications* 11 (1): 5843.
1032 https://doi.org/10.1038/s41467-020-19702-z.

1033 Luo, Liheng, Michael Gribskov, and Sufang Wang. 2022. "Bibliometric Review of ATAC-Seq and Its
1034 Application in Gene Expression." *Briefings in Bioinformatics*, March.
1035 https://doi.org/10.1093/BIB/BBAC061.

1036 Machlab, Dania, Lukas Burger, Charlotte Soneson, Filippo M. Rijli, Dirk Schübeler, and Michael B.
1037 Stadler. 2022. "MonaLisa: An R/Bioconductor Package for Identifying Regulatory Motifs."
1038 *Bioinformatics* 38 (9): 2624–25. https://doi.org/10.1093/BIOINFORMATICS/BTAC102.

1039 Mayer, Michael. 2023. "R Package ' SplitTools ': Tools for Data Splitting. R Package Version 1.0.1.
1040 Https://Cran.r-Project.Org/Web/Packages/SplitTools/Index.Html."

1041 Monaco, Gianni, Bernett Lee, Weili Xu, Seri Mustafah, You Yi Hwang, Christophe Carré, Nicolas
1042 Burdin, et al. 2019. "RNA-Seq Signatures Normalized by MRNA Abundance Allow Absolute
1043 Deconvolution of Human Immune Cell Types." *Cell Reports* 26 (6): 1627-1640.e7.

1044 https://doi.org/10.1016/J.CELREP.2019.01.041.

1045 Mumbach, Maxwell R., Ansuman T. Satpathy, Evan A. Boyle, Chao Dai, Benjamin G. Gowen, Seung
1046 Woo Cho, Michelle L. Nguyen, et al. 2017. "Enhancer Connectome in Primary Human Cells
1047 Identifies Target Genes of Disease-Associated DNA Elements." *Nature Genetics* 49 (11): 1602–
1048 12. https://doi.org/10.1038/ng.3963.

1049 Newman, Aaron M., Chih Long Liu, Michael R. Green, Andrew J. Gentles, Weiguo Feng, Yue Xu,
1050 Chuong D. Hoang, Maximilian Diehn, and Ash A. Alizadeh. 2015. "Robust Enumeration of Cell
1051 Subsets from Tissue Expression Profiles." *Nature Methods* 12 (5): 453–57.
1052 https://doi.org/10.1038/nmeth.3337.

1053 Newman, Aaron M., Chloé B. Steen, Chih Long Liu, Andrew J. Gentles, Aadel A. Chaudhuri, Florian
1054 Scherer, Michael S. Khodadoust, et al. 2019. "Determining Cell Type Abundance and Expression
1055 from Bulk Tissues with Digital Cytometry." *Nature Biotechnology* 37 (7): 773–82.
1056 https://doi.org/10.1038/s41587-019-0114-2.

1057 Peng, Xianlu Laura, Richard A. Moffitt, Robert J. Torphy, Keith E. Volmar, and Jen Jen Yeh. 2019. "De
1058 Novo Compartment Deconvolution and Weight Estimation of Tumor Samples Using DECODER."
1059 *Nature Communications* 10 (1): 4729. https://doi.org/10.1038/s41467-019-12517-7.

1060 Perez, Cristina, Cirino Botta, Aintzane Zabaleta, Noemi Puig, Maria-Teresa Cedena, Ibai Goicoechea,
1061 Daniel Alameda, et al. 2020. "Immunogenomic Identification and Characterization of
1062 Granulocytic Myeloid-Derived Suppressor Cells in Multiple Myeloma." *Blood* 136 (2): 199–209.
1063 https://doi.org/10.1182/BLOOD.2019004537.

1064 Qiu, Yixuan, Jiebiao Wang, Jing Lei, and Kathryn Roeder. 2021. "Identification of Cell-Type-Specific
1065 Marker Genes from Co-Expression Patterns in Tissue Samples." *Bioinformatics* 37 (19): 3228–34.
1066 https://doi.org/10.1093/BIOINFORMATICS/BTAB257.

1067 Racle, Julien, and David Gfeller. 2020. "EPIC: A Tool to Estimate the Proportions of Different Cell
1068 Types from Bulk Gene Expression Data." In *Methods in Molecular Biology*, 2120:233–48.
1069 Humana Press Inc. https://doi.org/10.1007/978-1-0716-0327-7_17.

1070 Racle, Julien, Kaat de Jonge, Petra Baumgaertner, Daniel E. Speiser, and David Gfeller. 2017.
1071 "Simultaneous Enumeration of Cancer and Immune Cell Types from Bulk Tumor Gene
1072 Expression Data." *ELife* 6 (November). https://doi.org/10.7554/eLife.26476.
1073 Ram-Mohan, Nikhil, Simone A Thair, Ulrike M Litzenburger, Steven Cogill, Nadya Andini, Xi Yang,
1074 Howard Y Chang, and Samuel Yang. 2021. "Profiling Chromatin Accessibility Responses in
1075 Human Neutrophils with Sensitive Pathogen Detection." *Life Science Alliance* 4 (8).
1076 https://doi.org/10.26508/LSA.202000976.
1077 Regner, Matthew J., Kamila Wisniewska, Susana Garcia-Recio, Aatish Thennavan, Raul Mendez-
1078 Giraldez, Venkat S. Malladi, Gabrielle Hawkins, et al. 2021. "A Multi-Omic Single-Cell Landscape
1079 of Human Gynecologic Malignancies." *Molecular Cell* 81 (23): 4924-4941.e10.
1080 https://doi.org/10.1016/j.molcel.2021.10.013.
1081 Risso, Davide, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. 2011. "GC-Content Normalization
1082 for RNA-Seq Data." *BMC Bioinformatics* 12 (December): 480. https://doi.org/10.1186/1471-
1083 2105-12-480/FIGURES/7.
1084 Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K.
1085 Smyth. 2015. "Limma Powers Differential Expression Analyses for RNA-Sequencing and
1086 Microarray Studies." *Nucleic Acids Research* 43 (7): e47. https://doi.org/10.1093/nar/gkv007.
1087 Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. "EdgeR: A Bioconductor Package
1088 for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1): 139–
1089 40. https://doi.org/10.1093/BIOINFORMATICS/BTP616.
1090 Rozowsky, Joel, Jiahao Gao, Beatrice Borsari, Roderic Guigó, Thomas R Gingeras, Mark Gerstein,
1091 Yucheng T Yang, et al. 2023. "The EN-TEx Resource of Multi-Tissue Personal Epigenomes &
1092 Variant-Impact Models." *Cell* 186 (7): 1493-1511.e40.
1093 https://doi.org/10.1016/j.cell.2023.02.018.
1094 Sanseviero, Emilio. 2019. "NK Cell-Fc Receptors Advance Tumor Immunotherapy." *Journal of Clinical*
1095 *Medicine* 8 (10): 1667. https://doi.org/10.3390/JCM8101667.

1096 Satpathy, Ansuman T., Jeffrey M. Granja, Kathryn E. Yost, Yanyan Qi, Francesca Meschi, Geoffrey P.
1097 McDermott, Brett N. Olsen, et al. 2019. "Massively Parallel Single-Cell Chromatin Landscapes of
1098 Human Immune Cell Development and Intratumoral T Cell Exhaustion." *Nature Biotechnology*
1099 37 (8): 925–36. https://doi.org/10.1038/s41587-019-0206-z.
1100 Smith, Jason P, M Ryan Corces, Jin Xu, Vincent P Reuter, Howard Y Chang, and Nathan C Sheffield.
1101 2021. "PEPATAC: An Optimized Pipeline for ATAC-Seq Data Analysis with Serial Alignments."
1102 *NAR Genomics and Bioinformatics* 3 (4). https://doi.org/10.1093/NARGAB/LQAB101.
1103 Stuart, Tim, Avi Srivastava, Shaista Madad, Caleb A. Lareau, and Rahul Satija. 2021. "Single-Cell
1104 Chromatin State Analysis with Signac." *Nature Methods* 18 (November): 1333–41.
1105 https://doi.org/10.1038/s41592-021-01282-5.
1106 Sturm, Gregor, Francesca Finotello, Florent Petitprez, Jitao David Zhang, Jan Baumbach, Wolf H
1107 Fridman, Markus List, and Tatsiana Aneichyk. 2019. "Comprehensive Evaluation of
1108 Transcriptome-Based Cell-Type Quantification Methods for Immuno-Oncology." *Bioinformatics*
1109 35 (14): i436–45. https://doi.org/10.1093/bioinformatics/btz363.
1110 Subramanian, Aravind, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael
1111 A Gillette, Amanda Paulovich, et al. 2005. "Gene Set Enrichment Analysis: A Knowledge-Based
1112 Approach for Interpreting Genome-Wide Expression Profiles." *PNAS* 102 (43): 15545–50.
1113 https://doi.org/10.1073/pnas.0506580102.
1114 Teschendorff, Andrew E., Tianyu Zhu, Charles E. Breeze, and Stephan Beck. 2020. "EPISCORE: Cell
1115 Type Deconvolution of Bulk Tissue DNA Methylomes from Single-Cell RNA-Seq Data." *Genome*
1116 *Biology* 21 (September). https://doi.org/10.1186/s13059-020-02126-9.
1117 The ENCODE Project Consortium, Jill E. Moore, Michael J. Purcaro, Henry E. Pratt, Charles B. Epstein,
1118 Noam Shoresh, and Jessika Adrian. 2020. "Expanded Encyclopaedias of DNA Elements in the
1119 Human and Mouse Genomes." *Nature* 583 (July): 699–710. https://doi.org/10.1038/s41586-
1120 020-2493-4.
1121 Trizzino, Marco, Avery Zucco, Sandra Deliard, Fang Wang, Elisa Barbieri, Filippo Veglia, Dmitry

1122 Gabrilovich, and Alessandro Gardini. 2021. “EGR1 Is a Gatekeeper of Inflammatory Enhancers in
1123 Human Macrophages.” *Science Advances* 7 (3).
1124 https://doi.org/10.1126/SCIADV.AAZ8836/SUPPL_FILE/AAZ8836_TABLE_S7.XLSX.
1125 Ucar, Duygu, Eladio J. Márquez, Cheng Han Chung, Radu Marches, Robert J. Rossi, Asli Uyar, Te Chia
1126 Wu, et al. 2017. “The Chromatin Accessibility Signature of Human Immune Aging Stems from
1127 CD8+ T Cells.” *Journal of Experimental Medicine* 214 (10): 3123–44.
1128 https://doi.org/10.1084/jem.20170416.
1129 Visser, Karin E. de, and Johanna A. Joyce. 2023. “The Evolving Tumor Microenvironment: From
1130 Cancer Initiation to Metastatic Outgrowth.” *Cancer Cell* 41 (3): 374–403.
1131 https://doi.org/10.1016/J.CCELL.2023.02.016.
1132 Watt, Stephen, Louella Vasquez, Klaudia Walter, Alice L. Mann, Kousik Kundu, Lu Chen, Ying Sims, et
1133 al. 2021. “Genetic Perturbation of PU.1 Binding and Chromatin Looping at Neutrophil Enhancers
1134 Associates with Autoimmune Disease.” *Nature Communications* 12 (April): 2298.
1135 https://doi.org/10.1038/S41467-021-22548-8.
1136 Welch, Ryan P., Chee Lee, Paul M. Imbriano, Snehal Patil, Terry E. Weymouth, R. Alex Smith, Laura J.
1137 Scott, and Maureen A. Sartor. 2014. “ChIP-Enrich: Gene Set Enrichment Testing for ChIP-Seq
1138 Data.” *Nucleic Acids Research* 42 (13): e105. https://doi.org/10.1093/NAR/GKU463.
1139 Xin, Jingxue, Hui Zhang, Yaoxi He, Zhana Duren, Caijuan Bai, Lang Chen, Xin Luo, et al. 2020.
1140 “Chromatin Accessibility Landscape and Regulatory Network of High-Altitude Hypoxia
1141 Adaptation.” *Nature Communications* 11 (1): 4928. https://doi.org/10.1038/S41467-020-18638-
1142 8.
1143 Yu, Guangchuang, Li Gen Wang, and Qing Yu He. 2015. “ChIPseeker: An R/Bioconductor Package for
1144 ChIP Peak Annotation, Comparison and Visualization.” *Bioinformatics* 31 (14): 2382–83.
1145 https://doi.org/10.1093/BIOINFORMATICS/BTV145.
1146 Zeng, Wanwen, Xi Chen, Zhana Duren, Yong Wang, Rui Jiang, and Wing Hung Wong. 2019. “DC3 Is a
1147 Method for Deconvolution and Coupled Clustering from Bulk and Single-Cell Genomics Data.”

1148 *Nature Communications* 10 (October): 4613. https://doi.org/10.1038/s41467-019-12547-1.

1149 Zhang, Hanyu, Ruoyi Cai, James Dai, and Wei Sun. 2021. "EMeth: An EM Algorithm for Cell Type

1150 Decomposition Based on DNA Methylation Data." *Scientific Reports* 11 (March): 5717.

1151 https://doi.org/10.1038/s41598-021-84864-9.

1152 Zhang, Kai, James D Hocker, Michael Miller, Xiaomeng Hou, Joshua Chiou, Olivier B Poirion, Yunjiang

1153 Qiu, et al. 2021. "A Single-Cell Atlas of Chromatin Accessibility in the Human Genome." *Cell* 184

1154 (24): 5985-6001.e19. https://doi.org/10.1016/j.cell.2021.10.024.

1155 Zhang, Ping, Harindra E. Amarasinghe, Justin P. Whalley, Chwen Tay, Hai Fang, Gabriele Migliorini,

1156 Andrew C. Brown, et al. 2022. "Epigenomic Analysis Reveals a Dynamic and Context-Specific

1157 Macrophage Enhancer Landscape Associated with Innate Immune Activation and Tolerance."

1158 *Genome Biology* 23 (1): 136. https://doi.org/10.1186/S13059-022-02702-1.

1159 Zhang, Yong, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein,

1160 Chad Nussbaum, et al. 2008. "Model-Based Analysis of ChIP-Seq (MACS)." *Genome Biology* 9

1161 (9): R137. https://doi.org/10.1186/GB-2008-9-9-R137/FIGURES/3.

1162

1163 **List of abbreviations:**

1164 **ATAC:** Assay for Transposase-Accessible chromatin

1165 **CBP(s):** chromatin binding protein(s)

1166 **ChIP-seq:** chromatin immunoprecipitation followed by sequencing

1167 **DC:** dendritic cells

1168 **NK:** natural killer cells

1169 **PCA:** principal component analysis

1170 **RMSE:** root mean squared error

1171 **TCGA :** The Cancer Genome Atlas

1172 **TF(s):** transcription factor(s)

1173 **TSS:** transcription start site

UMAP: Uniform Manifold Approximation

Main Figures.

Figure 1

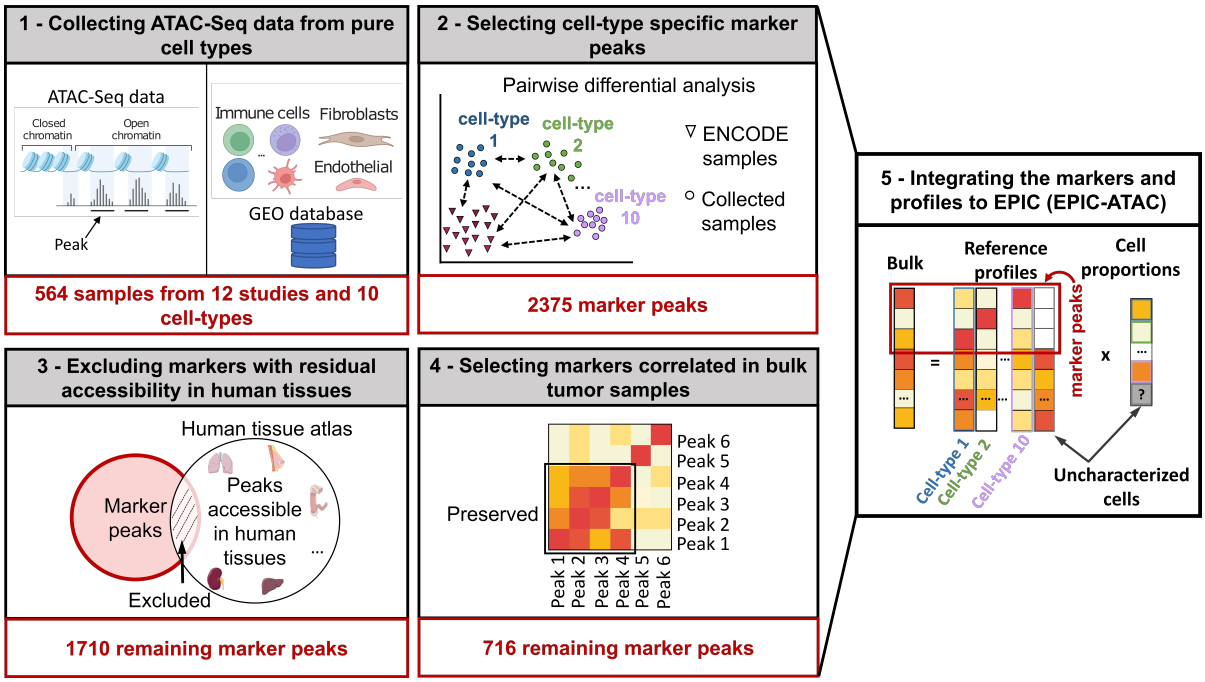


Figure 2

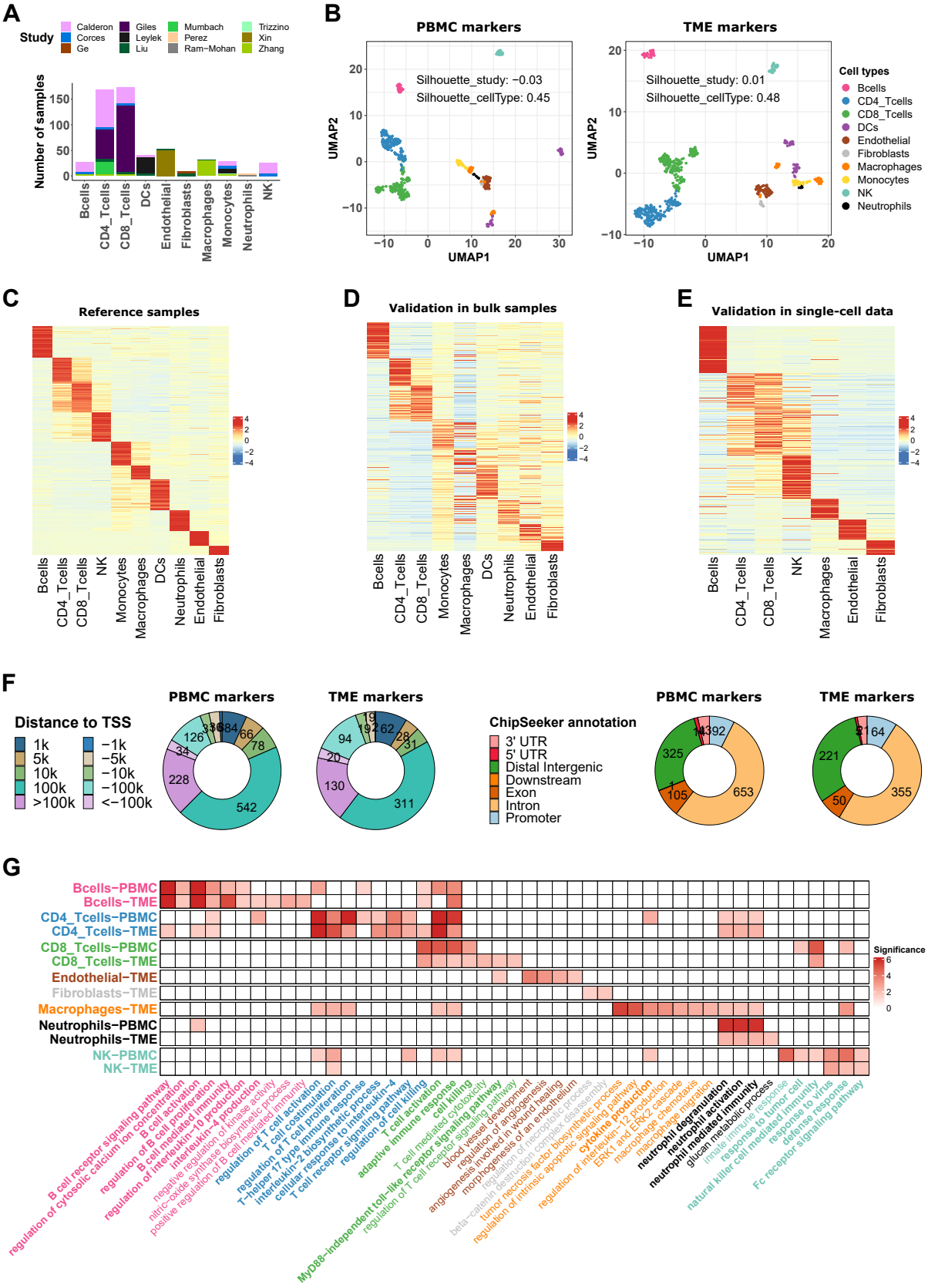

A
Study
Calderon
Corces
Ge
Giles
Leylek
Liu
Mumbach
Perez
Ram−Mohan
Trizzino
Xin
Zhang
Number of samples
Bcells
CD4_Tcells
CD8_Tcells
DCs
Endothelial
Fibroblasts
Macrophages
Monocytes
Neutrophils
NK
B
PBMC markers
Silhouette_study: −0.03
Silhouette_cellType: 0.45
TME markers
Silhouette_study: 0.01
Silhouette_cellType: 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

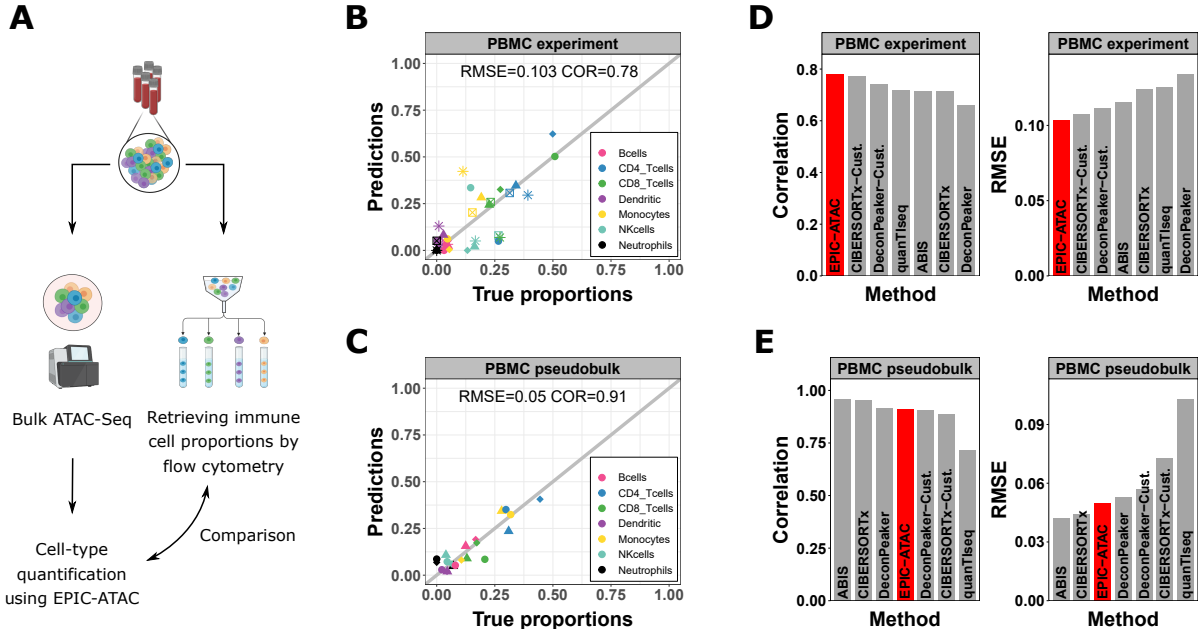

A
Bulk ATAC-Seq
Retrieving immune cell proportions by flow cytometry
Comparison
Cell-type quantification using EPIC-ATAC
B
PBMC experiment
RMSE=0.103 COR=0.78
Predictions
True proportions
Bcells
CD4_Tcells
CD8_Tcells
Dendritic
Monocytes
NKcells
Neutrophils
C
PBMC pseudobulk
RMSE=0.05 COR=0.91
Predictions
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

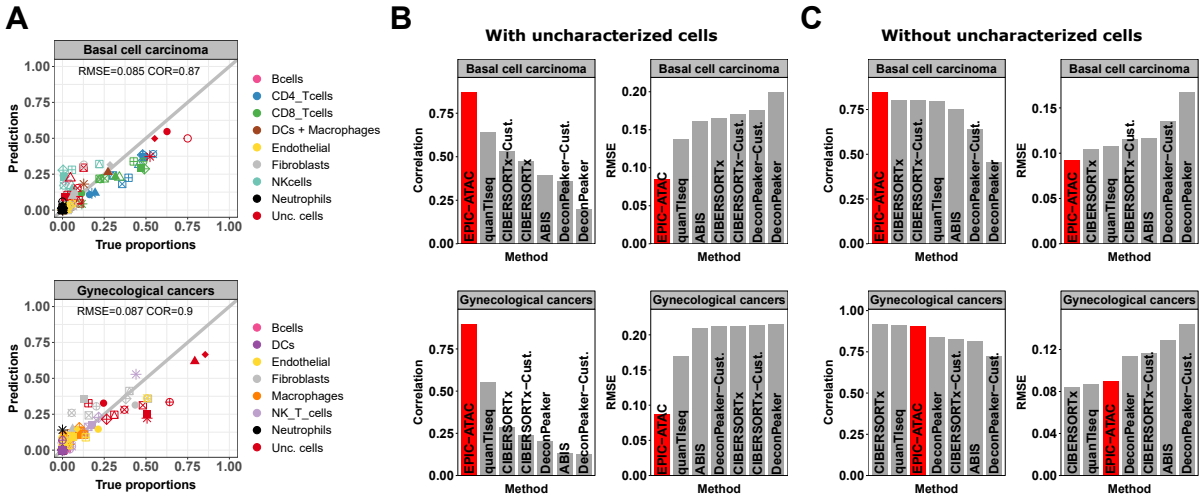

A
Basal cell carcinoma
RMSE=0.085 COR=0.87
Predictions
True proportions
Bcells
CD4_Tcells
CD8_Tcells
DCs + Macrophages
Endothelial
Fibroblasts
NKcells
Neutrophils
Unc. cells
Gynecological cancers
RMSE=0.087 COR=0.9
Bcells
DCs
Endothelial
Fibroblasts
Macrophages
NK_T_cells
Neutrophils
Unc. cells
B
With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

Figure 5

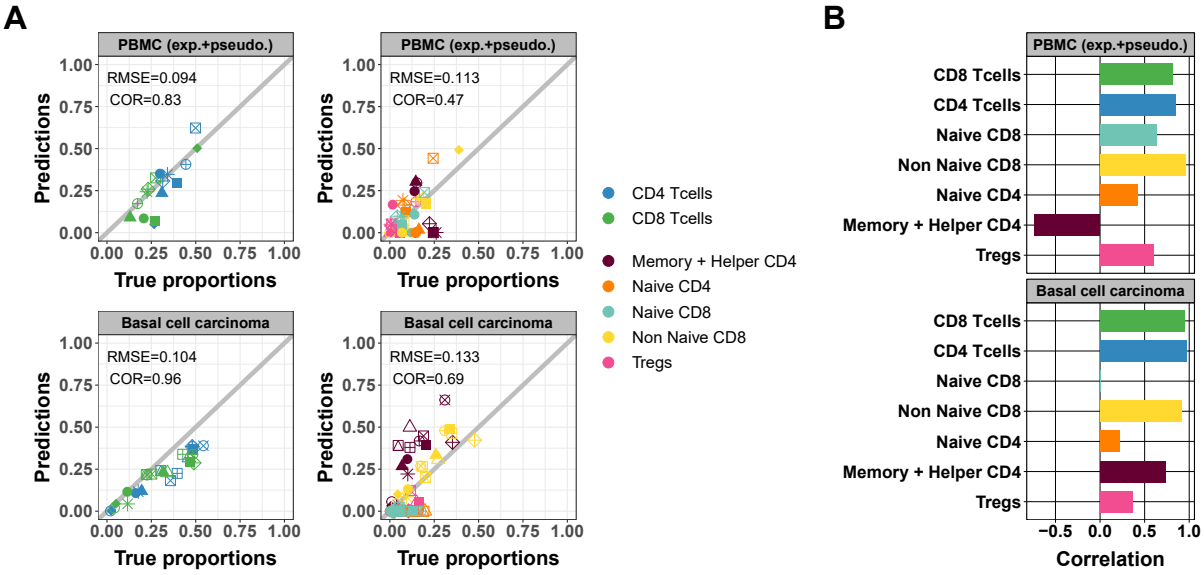

A
PBMC (exp.+pseudo.)
RMSE=0.094
COR=0.83
PBMC (exp.+pseudo.)
RMSE=0.113
COR=0.47
Basal cell carcinoma
RMSE=0.104
COR=0.96
Basal cell carcinoma
RMSE=0.133
COR=0.69
Predictions
True proportions
0.00
0.25
0.50
0.75
1.00
CD4 Tcells
CD8 Tcells
Memory + Helper CD4
Naive CD4
Naive CD8
Non Naive CD8
Tregs
B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

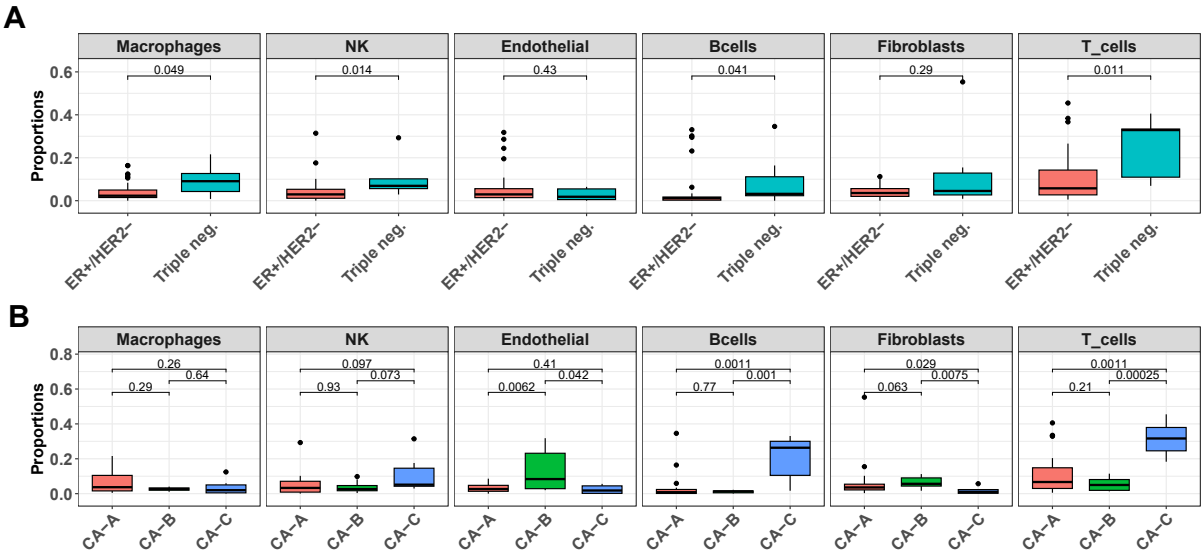
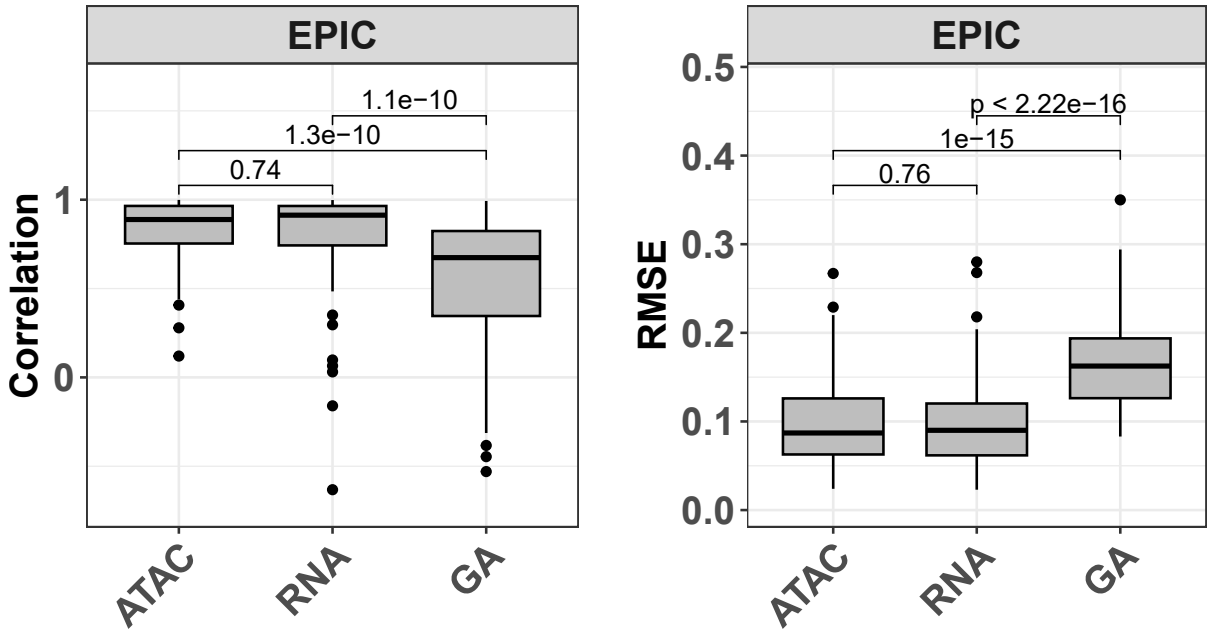

A
Macrophages
NK
Endothelial
Bcells
Fibroblasts
T_cells
Proportions
0.0
0.2
0.4
0.6
0.049
0.014
0.43
0.041
0.29
0.011
ER+/HER2−
Triple neg.
B
0.8
0.26
0.64
0.29
0.097
0.073
0.93
0.41
0.042
0.0062
0.0011
0.001
0.77
0.029
0.0075
0.063
0.0011
0.00025
0.21
CA−A
CA−B
CA−C

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