DEPARTMENT OF CANCER AND GENOMIC SCIENCES

COLLEGE OF MEDICINE AND HEALTH

UNIVERSITY OF BIRMINGHAM

Master Thesis

DECONVOLUTING LIQUID BIOPSY CELL-TYPE COMPOSITIONS FROM NUCLEOSOME FOOTPRINTS

Daniel Eigensatz

April 2025

Supervised by:

Dr. Roland Arnold

Department of Cancer and Genomic Sciences

University of Birmingham

Dr. Zsolt Balázs

KrauthammerLab

University of Zurich

Approved by: Dr. Roland Arnold Signature:  Date: 11/04/2025

Contents

Abstract

Circulating cell-free DNA (cfDNA) in liquid biopsies is recognized as a promising biomarker for early detection of cancers, guiding therapy, and monitoring drug resistance. However, cfDNA consists of fragmented DNA shed into the circulation and represents a heterogenous mixture derived from multiple cell types, making it challenging to trace its cellular origin. In this study, a novel approach is proposed that applies the EPIC-ATAC framework to deconvolute cell-type compositions from cfDNA based on nucleosome footprints, using ATAC-Seq cell-type specific chromatin accessibility markers. Validation on synthetic mixtures confirmed the accurate deconvolution at low sequencing depths. When applied to cfDNA of healthy and cancer patients, the estimated contributions of major immune cell types, such as neutrophils and lymphocytes seem biologically plausible. However, other cell types were overestimated, highlighting the need to refine the marker specificity to improve the performance of the cfDNA deconvolution.

1 Introduction

Detecting and diagnosing cancer at an early stage is critical for improving patient survival and treatment outcomes[1–3]. In this context, cfDNA in liquid biopsies has emerged as a potential biomarker as it provides a snapshot of the genetic and epigenetic landscape of its cell of origin[4]. cfDNA is primarily released into circulation through cellular processes such as apoptosis and necrosis. It has a short half-life of 15 minutes to approximately 2.5 hours[5]. The majority of cfDNA in blood plasma of healthy individuals originates from hematopoietic cells due to their frequent turnover[6]. The levels and composition can be altered in cancer, where circulating tumor DNA (ctDNA) is a significant fraction of the total cfDNA[7].

The fragmentation profile is one of the most informative features of cfDNA. Apoptotic cells undergo DNA fragmentation in multiples of the nucleosome repeat length, resulting in distinct fragment sizes corresponding to mono-, di-, and trinucleosomes. This laddering pattern occurs

because apoptotic endonucleases cleave DNA in the linker region that connects the nucleosomal units. In other words, the cfDNA fragments can be traced to where the nucleosomes were wrapped around the DNA. Since the positioning of nucleosomes is influenced by the cell-type specific organization of chromatin, the nucleosome footprints observed in cfDNA can be used to infer the tissue and cell type of origin[8].

Unlike previous cfDNA deconvolution studies that relied on DNA methylation to infer the cell-type composition[9,10], this work explores a novel approach based on chromatin accessibility and nucleosome footprints. ATAC-Seq (Assay for Transposase-Accessible Chromatin using sequencing) is a method for profiling genome-wide chromatin accessibility, providing insights into regulatory elements and cell-type specific chromatin organization. Based on this principle, the EPIC-ATAC framework developed by Gabriel et al.[11] is originally designed to deconvolute bulk ATAC-Seq samples. By mapping nucleosome footprints in cfDNA to the ATAC-Seq derived marker from the EPIC-ATAC framework[11], this study aims to evaluate whether the chromatin accessibility derived signals can be used to infer the cell-type composition in cfDNA through nucleosome footprints, offering an alternative to the DNA methylation based deconvolution approaches.

2 Results

2.1 Evaluation of cell-type specific markers using ATAC-Seq data

For accurate deconvolution reliable cell-type specific markers and reference profiles are critical[11]. To assess the specificity of the 716 marker regions identified by Gabriel et al.[11] for 9 cell types (B cells, CD4+ T cells, CD8+ T cells, natural killer (NK) cells, macrophages, dendritic cells (DCs), neutrophils, endothelial cells, and fibroblasts), ATAC-Seq data from reference and tissue samples, as well as single-cell ATAC-Seq data were analyzed.

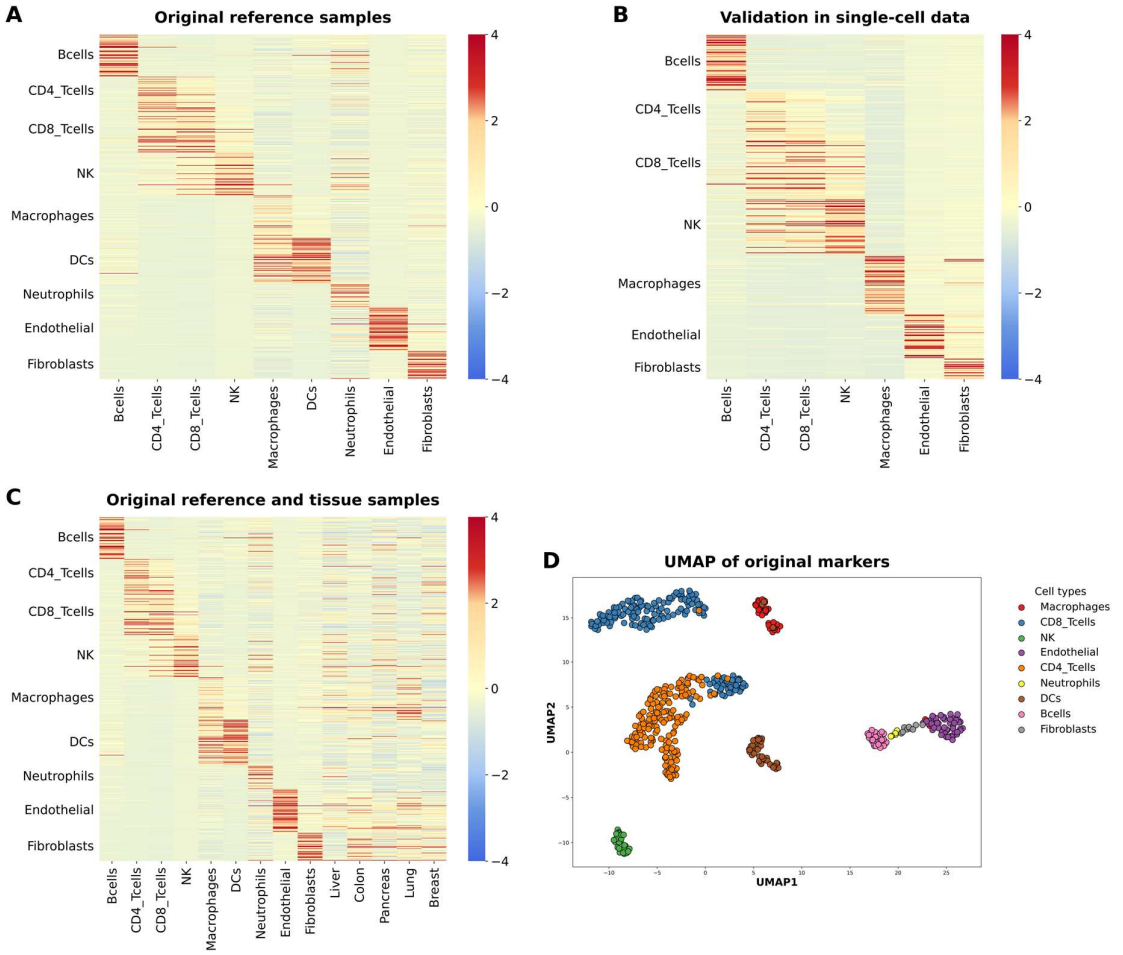

Figure 1: **A)** Scaled chromatin accessibility of the original markers (rows) in each cell type (columns) in ATAC-Seq reference samples used to identify the markers. **B)** Scaled chromatin accessibility of the original markers in the scATAC-Seq dataset[12]. **C)** Scaled chromatin accessibility of each original marker (rows) in the reference samples and in the ENCODE samples from diverse tissues (columns). **D)** UMAP clustering based on the original markers. The colors represent different cell types.

The analysis aimed to validate whether the chromatin accessibility in the markers is unique to the cell type and whether this is consistent across datasets. As previously reported by Gabriel et al.[11] and shown in Figure 1A and 1B, the markers for endothelial cells and fibroblasts show good specificity with minimal cross-marker signals. However, when analyzing individual reference samples instead of averaging multiple reference samples as performed by Gabriel et al, significant cross-marker signal overlap can be observed in immune cells. In particular, CD4+ and CD8+ T cells, NK cells, neutrophils and B cells show accessibility in multiple markers. Similarly, macrophages and DCs share accessibility in several markers. A comparable trend is observed in the tissue samples (Figure 1C), where markers show overlapping

accessibility. Although residual overlapping regions were filtered out using a human tissue atlas[12] in their differential accessibility analyis[11], some shared accessibility remains when looking at individual tissue reference samples. In the UMAP projection (Figure 1D) the markers primarily cluster by cell type rather than by study of origin. However, fibroblasts and neutrophils cluster together which suggests a degree of similarity in their chromatin accessibility.

2.2 Evaluation of new cell-type specific markers using ATAC-Seq data

The 716 markers identified by Gabriel et al. were extended by 319 additional markers through pairwise differential accessibility analysis (See methods). The new markers are derived from five cell and tissue types: Hepatocytes, breast carcinoma (BRCA), colon adenocarcinoma (COAD), lung adenocarcinoma (LUAD), and lung squamous cell carcinoma (LUSC). The extension resulted in a total of 1,035 markers covering 14 distinct cell and tissue types. The specificity of the newly identified markers was evaluated in ATAC-Seq reference, single-cell, and tissue datasets. As shown in Figure 2A, hepatocyte markers indicate cross-marker signals in macrophages, neutrophils, and fibroblasts. Compared to bulk ATAC-Seq, the single-cell data (Figure 2B) show improved specificity for hepatocyte markers, but cross-marker accessibility remains present in fibroblasts. A similar trend is observed for the tissue samples (Figure 2C), suggesting that the hepatocyte markers may not be completely cell-type specific despite filtered out the regions that show accessibility in universally accessible modules of human tissues and other cell types[12] (see Methods). In contrast, the selected cancer specific markers (BRCA, COAD, LUAD, and LUSC) appear to have high specificity in the reference samples (Figure 2A) and they do not show cross-marker signal across other cell types. In tissue samples (Figure 2C), BRCA markers show accessibility in breast tissue, and LUAD markers in lung tissue, further supporting their specific chromatin accessibility profiles. The UMAP projection (Figure 2D) illustrates the clustering of the cell types based on the extended markers.

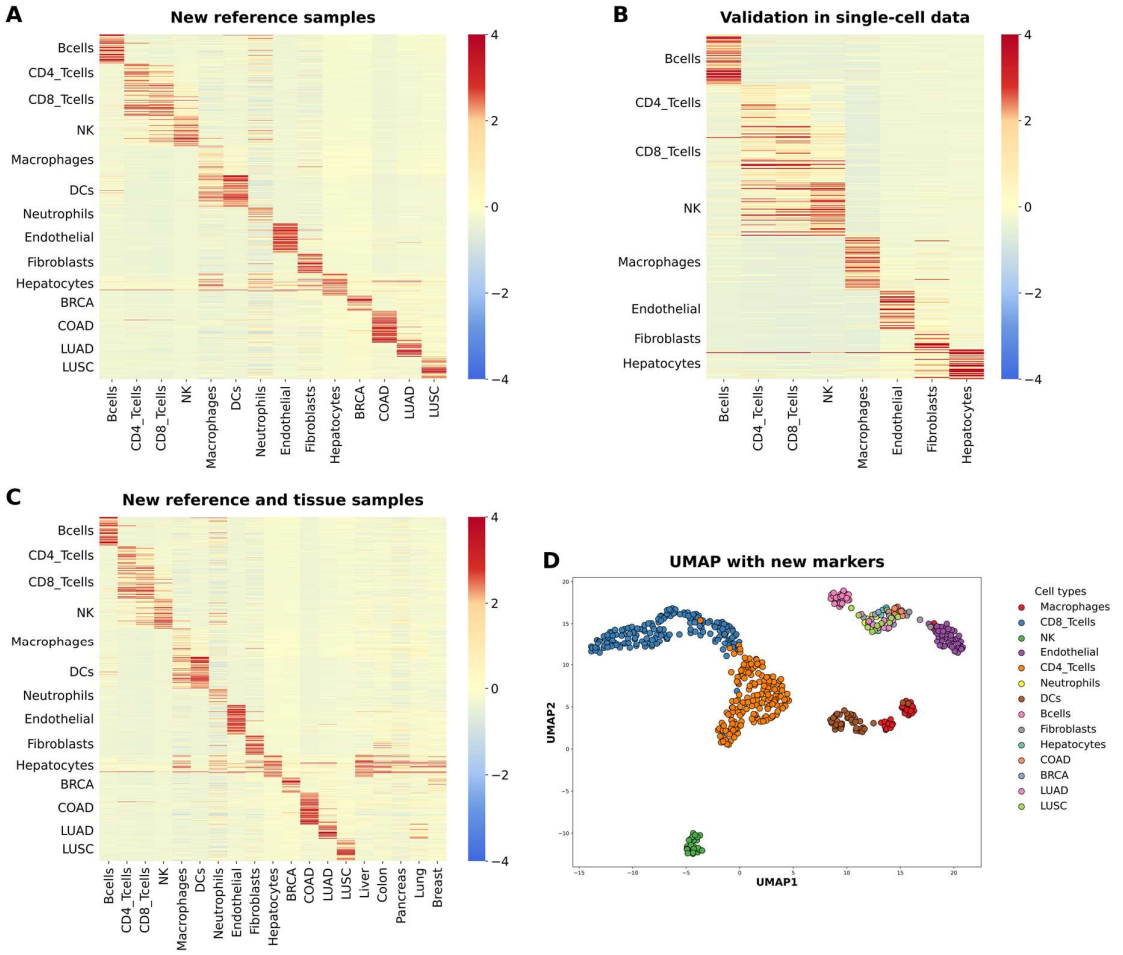

Figure 2: **A)** Scaled chromatin accessibility of the extended markers (rows) in each cell type (columns) in ATAC-Seq reference samples used to identify the markers. **B)** Scaled chromatin accessibility of the extended markers in the scATAC-Seq dataset[12]. **C)** Scaled chromatin accessibility of the extended markers (rows) in the reference samples and in the ENCODE samples from diverse tissues (columns). **D)** UMAP clustering based on the extended markers. The colors represent different cell types.

While most cell types form distinct clusters, the newly identified markers cluster close with neutrophils and fibroblasts, showing no clear separation. This suggests common chromatin accessibility patterns between these cell types, which may influence the performance of the deconvolution.

2.3 Validation of the EPIC-ATAC deconvolution using synthetic mixtures

To evaluate the performance of the EPIC-ATAC algorithm in estimating cell-type proportions, synthetic mixtures with varying proportions (0%, 1%, 2%, 4%, 8%, 13%, 17%, 21% and 34%) were created (see Methods). The mixtures were based on the 716 cell-type markers extracted

from the reference profile of the 9 cell types. The samples did not contain noise or bias, meaning that all regions were preserved, and no modification or deletion was introduced. Since the markers were defined by the reference profile, the algorithm is expected to achieve near-perfect deconvolution in this scenario. The results are summarized in Figure 3 which shows scatterplots of true versus predicted proportions for each of the 9 cell types. Overall, the results demonstrated high concordance correlation coefficients (CCC) ranging from 0.766 to 0.993 and low root mean square error (RMSE) ranging from 0.013 to 0.098. In particular, macrophages (CCC = 0.993, RMSE = 0.013), endothelial cells (CCC = 0.989, RMSE = 0.016), and NK cells (CCC = 0.988, RMSE = 0.016) showed near-perfect CCC with low RMSE. B cells, CD8+ T cells, fibroblasts, and neutrophils also showed CCC values exceeding 0.93 and RMSE values below 0.04. However, CD4+ T cells (CCC = 0.766, RMSE = 0.068) and DCs (CCC = 0.788, RMSE = 0.098) had slightly weaker performance, suggesting minor deviation in the predicted proportions.

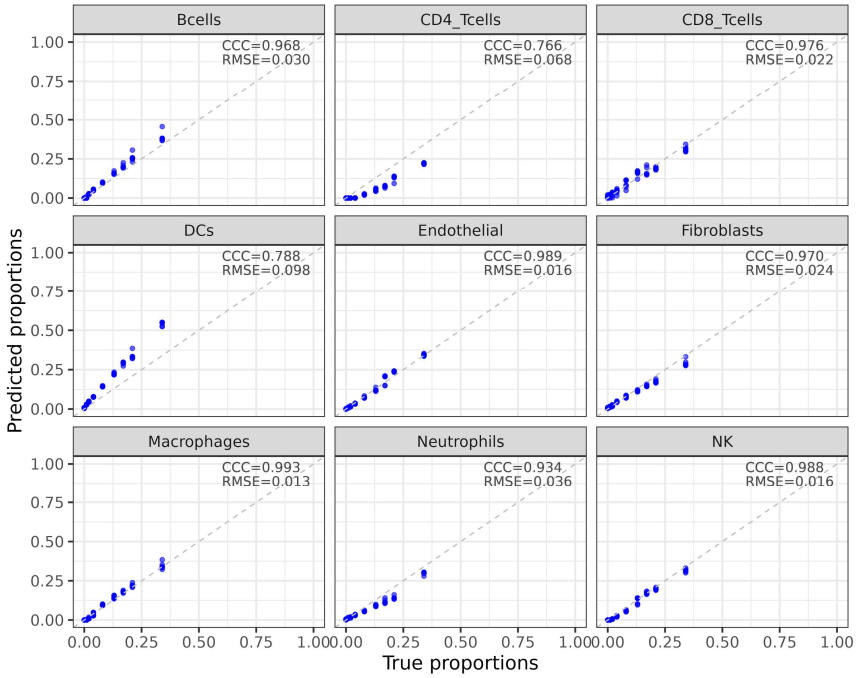

Figure 3: True versus predicted cell-type proportions after deconvolution for all cell types. The true proportions (x-axis) show the proportions assigned to the cell type in the synthetic sample. The predicted proportions (y-axis) show the estimation after deconvolution.

2.4 Impact of sequencing depth on the EPIC-ATAC performance

Since cfDNA samples have significantly lower coverage (2x – 8x) compared to the synthetic samples, the deconvolution was simulated under similar conditions. Down-sampling was performed by systematically reducing the sequencing depth (0.1x, 0.3x, 1x, 3x, 9x, 30x, 80x, and 245x) in the synthetic samples (see Methods). As shown in Figure 4, the deconvolution performance remained stable at sequencing depth of 3x with a high CCC (ranging from 0.786 to 0.991) and low RMSE (ranging from 0.015 to 0.102) values comparable to the results obtained in Figure 3. As the sequencing depth decreases below 3x, the deconvolution performance started to decline, where DCs (CCC = 0.741 and RMSE = 0.122) and CD4+ T cells (CCC = 0.736 and RSME = 0.076) show the highest variability among the cell types. At 0.1x coverage, the deconvolution performance further declined, with low CCC (ranging from 0.199 to 0.601) and high RMSE (ranging from 0.088 to 0.340), indicating a near-random distribution of predicted proportions. The results demonstrate that the EPIC-ATAC algorithm remains robust at 3x – 9x coverage, suggesting reliable cell-type proportions estimates can be obtained within the expected sequencing depth of the cfDNA samples.

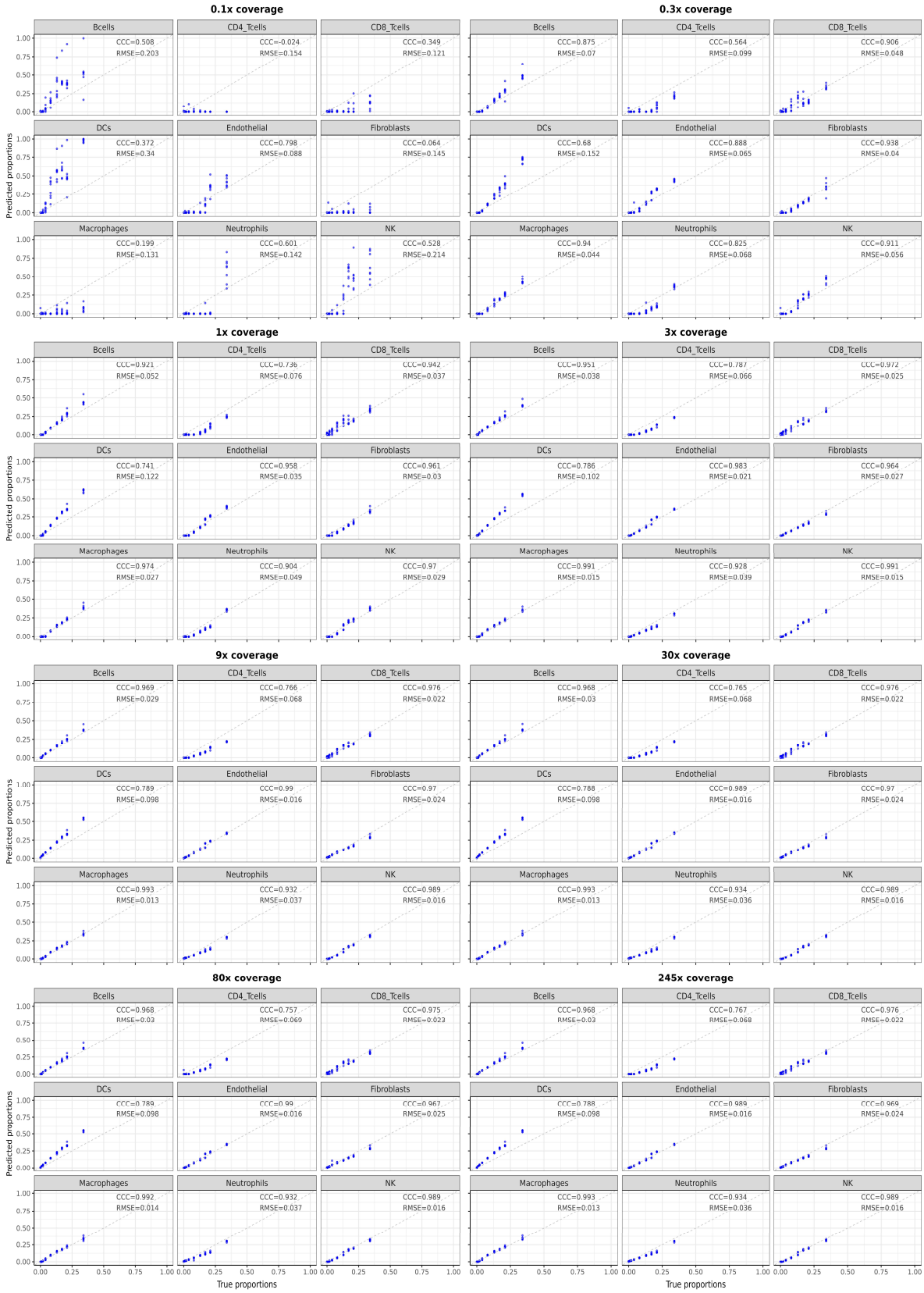

Figure 4: True versus predicted cell-type proportions across different sequencing depths, ranging from 0.1x to 245x coverage. Each subplot corresponds to the true (x-axis) and predicted (y-axis) proportions of the specific cell type at the given coverage.

2.5 Cell-type specific nucleosome footprints in cfDNA-Seq data

The nucleosome footprints were explored by smoothed z-score normalized cfDNA fragment center counts (see Methods) across universal open and closed chromatin regions, as well as the 716 marker regions of the 9 cell types. The profiles are visualized in Figure 5, which shows the averaged cfDNA fragment center distribution of 5 healthy individuals across 2 kb windows with shaded areas as the standard deviation across the samples.

The fragment center counts in universal open chromatin regions are lower in the center of the 2 kb window forming a characteristic U-shaped profile. This decrease suggests an increased nucleosome-free fragmentation, where cfDNA is more susceptible to degradation due to reduced nucleosome protection, resulting in fewer fragments accumulating at the center[8]. In contrast, the fragment center counts in universal closed chromatin regions are increased in the center of the 2 kb window. This pattern reflects the influence of nucleosome positioning on cfDNA fragmentation. Nucleosome-bound DNA is protected from cleavage, while fragmentation preferentially occurs at linker DNA regions between nucleosomes[8]. As a results, more cfDNA fragments are derived from nucleosome-bound regions with increased fragment center counts at these sites.

Since cfDNA originates from a heterogeneous mixture of cell types, the hypothesis was that if the marker regions are truly cell-type specific, the fragment center counts for the corresponding cell type should be decreased (indicating an increased contribution to open chromatin) in the mixture, while all other cell types should show increased fragment center counts (indicating an increased contribution to closed chromatin). The observed signal is the superimposed contribution of all cell types present in the sample. The expectation was that if a particular cell type was present in the sample, the superimposed cfDNA fragment signal would follow a U-shaped distribution similar to that observed for universal open chromatin regions.

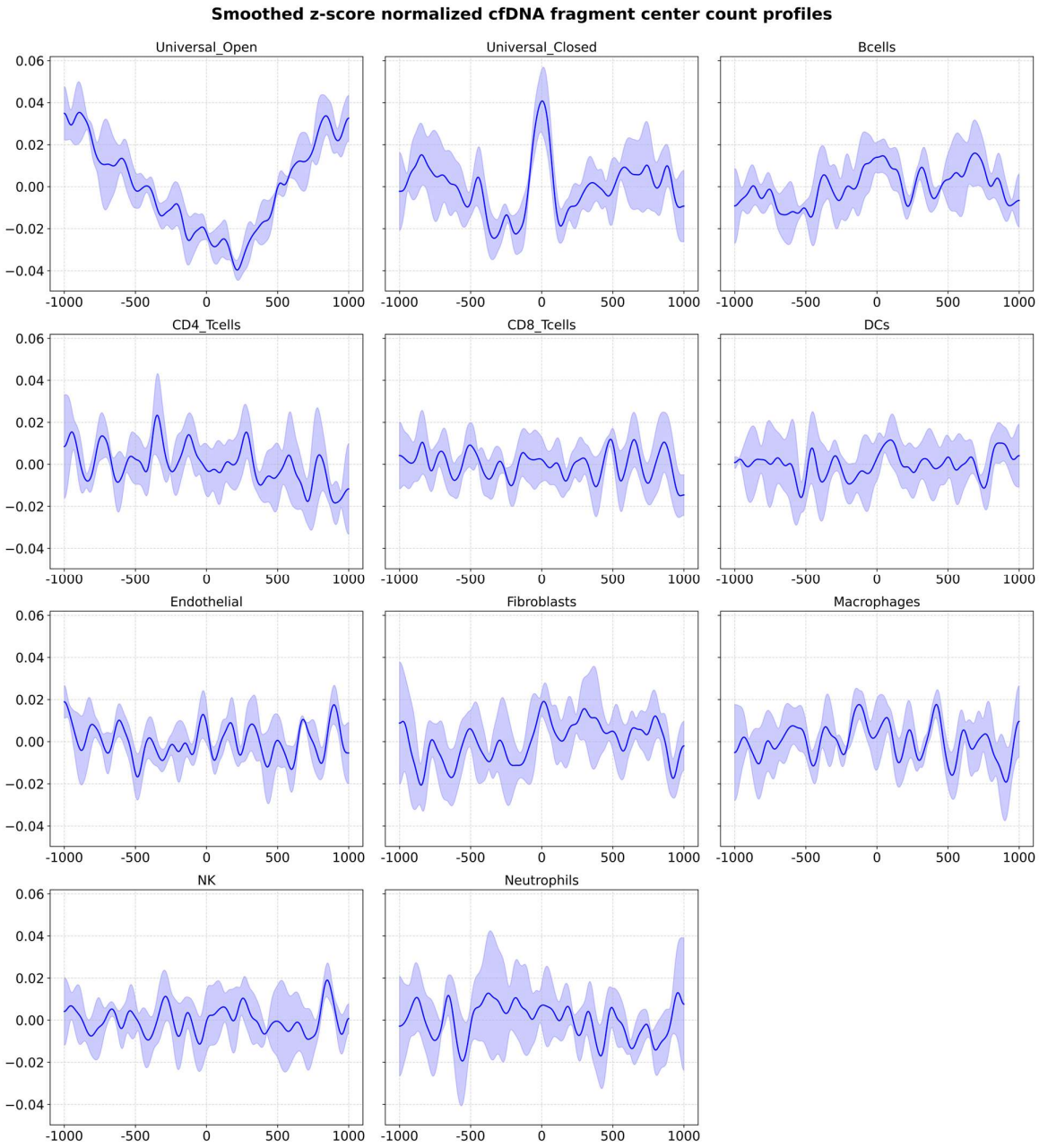

Figure 5: Smoothed z-score normalized cfDNA fragment center count profiles across universal open and closed chromatin regions and the cell-type specific markers, averaged over 5 healthy cfDNA samples. The shaded regions show the standard deviation across the samples.

Since neutrophils are known to be the dominant contributors to cfDNA in healthy individual[13], their markers were expected to show the strongest U-shaped pattern. However, while Figure 5 shows some degree of variability in the fragment center counts between the different cell types, the observed differences are less pronounced than expected. For this reason, only the most informative regions within a 100 bp window around the center of each marker were extracted

for the deconvolution. By focusing on this window size, it was assumed that the deconvolution algorithm can capture the variation between the cell types.

2.6 EPIC-ATAC deconvolution on real-world cfDNA samples

EPIC-ATAC was first applied to cfDNA samples from 5 healthy individuals. Deconvolution was performed using the 716 cell-type markers published by Gabriel et al.[11]. The predicted cell-type proportions are presented in Figure 6 with the highest median proportions observed for neutrophils (21.0%) and fibroblasts (20.6%), followed by CD4+ (11.6%) T cells and DCs (11.2%). Lower proportions were estimated for CD8+ (7.0%) T cells, macrophages (8.2%), B cells (7.3%), NK cells (6.8%) and endothelial cells (6.3%).

Among the cfDNA samples from the 61 cancer patients (Figure 6) the highest proportions were observed for neutrophils (19.4–21.6%) and fibroblasts (18.7–19.7%), followed by DCs (11.3–12.2%) and CD4+ T cells (9.4–12.3%). For macrophages (7.9–9.3%), NK cells (6.4–7.7%), B cells (7.1–8.2%), endothelial cells (6.5–7.6%), and CD8+ T cells (4.8–8.3%) lower proportions were predicted. The corresponding median proportions for all samples are summarized in Table 1.

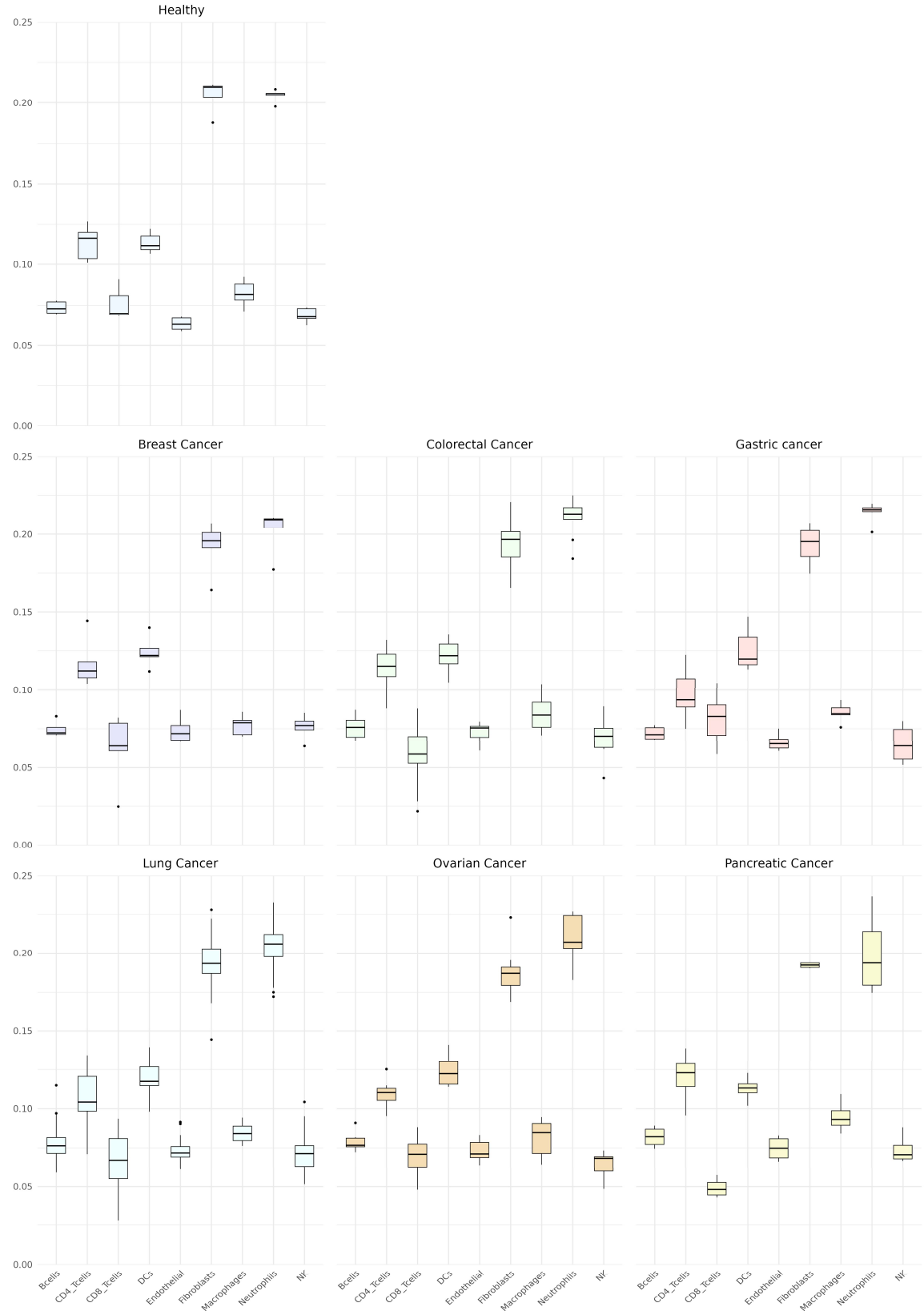

Figure 6: Estimated proportions (y-axis) of the cell types (x-axis) in the 5 healthy and 61 cfDNA cancer samples using the original markers for deconvolution.

	Healthy	Breast Cancer	Colorectal Cancer	Gastric Cancer	Lung Cancer	Ovarian Cancer	Pancreatic Cancer
Neutrophils	21.0	20.9	21.3	21.6	20.6	20.7	19.4
Fibroblasts	20.6	19.6	19.7	19.5	19.3	18.7	19.2
DCs	11.2	12.2	12.2	12.0	11.7	12.2	11.3
CD4+ T cells	11.6	11.2	11.5	9.4	10.4	11.0	12.3
Macrophages	8.2	7.9	8.4	8.5	8.4	8.5	9.3
NK cells	6.8	7.7	7.0	6.4	7.2	6.9	7.1
B cells	7.3	7.2	7.6	7.1	7.7	7.7	8.2
Endothelial cells	6.3	7.2	7.6	6.5	7.2	7.1	7.5
CD8+ T cells	7.0	6.4	5.9	8.3	6.7	7.1	4.8

Table 1: Median cell-type proportions in percentages across the 5 healthy and the 61 cancer cfDNA samples using the original markers for deconvolution.

2.7 EPIC-ATAC deconvolution using the extended markers

To evaluate whether additional cell and cancer tissue types (hepatocytes, BRCA, COAD, LUAD, and LUSC) can be detected, EPIC-ATAC was applied to the 5 healthy and 61 cancer cfDNA samples using the extended set of 1,035 marker regions (see Methods). Among the healthy cfDNA samples (Figure 7), the highest proportions were estimated for LUSC (21.3%), neutrophils (15.6%), LUAD (14.1% and fibroblasts (13.7%), and, followed by BRCA (7.2%), DCs (6.9%), CD4+ T cells (5.7%), and NK cells (3.4%). Lower proportions were observed for CD8+ T cells (2.8%), endothelial cells (2.2%), COAD (1.7%), hepatocytes (1.1%), macrophages (1.1%), and B cells (0.5%). Across the 61 cfDNA cancer samples (Figure 7), the highest proportions were predicted for LUSC (20.6–22.7%), LUAD (14.0–16.7%), and neutrophils (15.0–15.9%). Moderate contributions were observed for fibroblasts (12.3–13.0%), BRCA (7.9–9.1%), and DCs (6.5–7.1%).

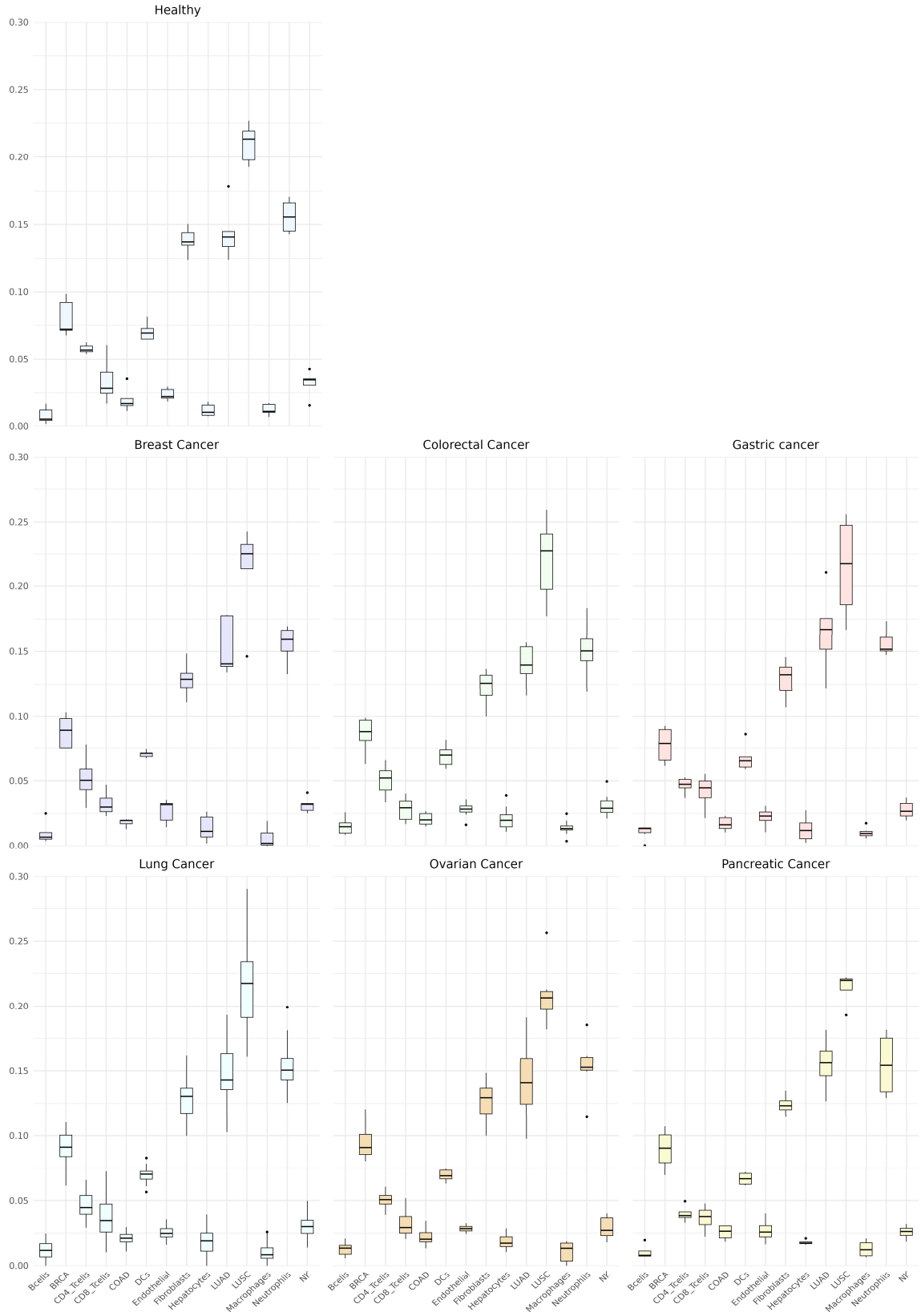

Figure 7: Estimated proportions (y-axis) of the cell types (x-axis) in the 5 healthy and 61 cfDNA cancer samples using the extended markers for deconvolution.

Lower proportions were estimated for CD4+ T cells (3.9–5.1%), CD8+ T cells (2.9–4.5%), endothelial cells (2.3–3.2%), NK cells (2.6–3.2%), COAD (1.6–2.7%), and hepatocytes (1.1–2.0%). B cells and macrophages were below 1% in all cancer samples. The corresponding median proportions for all samples are summarized in Table 2.

	Healthy	Breast Cancer	Colorectal Cancer	Gastric Cancer	Lung Cancer	Ovarian Cancer	Pancreatic Cancer
LUSC	21.3	22.5	22.7	21.8	21.7	20.6	22.0
Neutrophils	15.6	15.9	15.0	15.2	15.1	15.3	15.4
LUAD	14.1	14.1	14.0	16.7	14.3	14.1	15.6
Fibroblasts	13.7	12.9	12.6	13.2	13.0	12.9	12.3
BRCA	7.2	8.9	8.8	7.9	9.1	9.1	9.1
DCs	6.9	7.1	7.0	6.5	7.0	6.9	6.7
CD4+ T cells	5.7	5.0	5.2	4.7	4.5	5.1	3.9
NK cells	3.4	3.2	2.9	2.7	3.0	2.7	2.6
Endothelial cells	2.2	3.2	2.8	2.3	2.5	2.9	2.6
CD8+ T cells	2.8	3.0	2.9	4.5	3.5	2.9	3.8
COAD	1.7	2.0	2.0	1.6	2.1	2.0	2.7
Hepatocytes	1.1	1.1	2.0	1.2	1.9	1.7	1.7
B cells	0.5	0.7	0.1	0.1	0.1	0.1	0.1
Macrophages	1.1	0.0	0.1	0.1	0.1	0.1	0.1

Table 2: Median cell-type proportions in percentages across the 5 healthy and 61 cancer cfDNA samples using the extended markers for deconvolution.

3 Discussion

In this study, cfDNA nucleosome footprints within the cell-type specific ATAC-Seq markers are integrated into the EPIC-ATAC framework to perform deconvolution. The work presents an alternative approach to previous DNA methylation based deconvolution approaches[9,10]. In the context of cancer, changes of major immune, stromal and vascular cell type contributions in cfDNA may reflect immune infiltration, tumor burden, and treatment response[14]. Since cfDNA samples can be collected non-invasively, this approach has potential for early cancer

detection, when tumor derived signals are observed with characteristic changes in the tumor micro-environment.

To evaluate the performance of the EPIC-ATAC framework, synthetic mixture experiments validated the robust and accurate deconvolution down to 3x sequencing depth, which is comparable to the coverage of the cfDNA samples from Cristiano et al.[15].

The analysis of the cell-type specific nucleosome footprints within the ATAC-Seq markers did not confirm the initial hypothesis of decreased fragment center counts at the region center and the characteristic U-shaped profile was not observed across the cell types. Several factors may have contributed to this observation. First, the cfDNA signal was averaged across the markers of each cell type in Figure 5. Second, the markers may not be entirely cell-type specific, as suggested by the overlapping chromatin accessibility observed in both the reference and tissue samples (Figure 1). The residual overlap potentially limits the resolution of the nucleosome footprints for individual cell types. As a result, only the most informative segment, defined as the 100 bp window within each marker, was extracted and used for downstream deconvolution. Despite these limitations, the EPIC-ATAC was applied to real-world cfDNA samples from both healthy and cancer patients. Previous publications from Moss et al.[9] and Fox-Fisher et al.[16] reported that neutrophils (~32%) are the predominant source of cfDNA, followed by monocytes/macrophages (~11%), endothelial cells (~10%) and the lymphoid subtypes such as CD4+ T cells, CD8+ T cells, B cells, and NK cells with a combined contribution of ~30%. The contributions to fibroblasts and DCs are negligible in cfDNA. In this study, neutrophils were also identified as the most abundant cell type, and the lymphoid subtypes, macrophages, and endothelial cells fell approximately within the expected range, suggesting that EPIC-ATAC can detect major immune cell type contributions in cfDNA-Seq data. However, fibroblasts and DCs were predicted with unreasonably high proportions. As indicated by the UMAP clustering (Figure 1D), fibroblasts cluster close with neutrophils, and DCs partially with macrophages.

The proximity suggests similarity in their chromatin accessibility profiles, which is unfavourable for deconvolution and might lead to an artificially inflated predictions for fibroblasts and DCs.

To investigate whether EPIC-ATAC could detect hepatocytes, BRCA, COAD, LUAD, and LUSC signals, additional 319 markers were derived through pairwise differential accessibility analysis. Despite showing promising specificity in the reference and tissue samples (Figure 2A and 2C), UMAP clustering (Figure 2D) indicated limited separation between the new cell types, fibroblasts, and neutrophils. When applied to EPIC-ATAC, the results yielded in implausible high proportions for cancer-derived signals even in healthy individuals. Similarly, in cancer patient samples, the estimated proportions exceed biologically reasonable levels and did not reflect the corresponding tumor type. A potential explanation for the results could be the limitation imposed by the raw count matrix from Gabriel et al.[11] which contains 160,276 predefined genomic regions identified through differential analysis of the 9 cell types in their study. The raw count matrix was used in this study to obtain new markers through pairwise differential accessibility analysis. As a result, the marker selection of hepatocytes and cancer-specific tissues was constrained to these regions, where many contain residual accessibility in immune and stromal cells. This overlap likely reduces the specificity of the new markers and contributes to the biologically implausible estimates in both healthy and cancer cfDNA samples.

Future work could focus on developing a more comprehensive and unbiased reference atlas to improve the marker selection through differential accessibility analysis. These markers should be truly cell-type specific without overlapping chromatin accessibility in other cell types or tissues, which could result in more distinct nucleosome footprint profile and improve the performance of the cfDNA deconvolution to study the tumor micro-environment and detecting cancers.

4 Methods

4.1 Datasets

16 pure ATAC-Seq bulk samples from 9 studies were collected including B cells[17], CD4+ T cells[17], CD8+ T cells[17], NK cells[17], macrophages[18], DCs[19], neutrophils[20], endothelial cells[21], fibroblasts[22], and hepatocytes[23–25]. In addition, 5 distinct organ tissue samples were obtained from ENCODE[26] and include liver, colon, pancreas, lung, and breast. A raw count matrix of 564 pre-processed pure ATAC-Seq bulk samples from 12 studies was provided by Gabriel et al.[11] and includes the same immune, stromal and vascular cell types as previously mentioned except for hepatocytes. bigWig-formatted ATAC-Seq bulk samples from primary human tumors, including BRCA, COAD, LUAD, and LUSC, consisted of 10 samples per cancer type and were prepared by Corces et al.[27]. 66 cfDNA samples, including 5 healthy and 61 diseased cases (breast, colorectal, gastric, lung, ovarian, and pancreatic cancer), were provided by Cristiano et al.[15].

4.2 Pre-processing of raw ATAC-Seq datasets

Raw sequencing data (FASTQ files) were obtained from GEO using the SRA toolkit. Sequencing adapters were assessed with fastp[28], followed by trimming and removal of low-quality bases using Trimmomatic[29]. The pre-processed reads were aligned on the human genome (hg38) using bowtie2[30] with parameters: --very-sensitive -X 2000. Duplicates were removed applying SAMtools markdup[31] to minimize bias from over-amplified reads. After alignment, reads mapped to the ENCODE hg38 blacklist[32] and chromosome M were excluded. Read counts were obtained for 160,276 genomic regions that matched those used in the raw count matrix of Gabriel et al.[11] which allowed for the direct integration of the samples for the downstream analysis.

4.3 Pre-processing of bigWig-formatted ATAC-Seq datasets

Due to access restrictions, the ATAC-Seq datasets of primary human tumors were obtained in bigWig format rather than as raw sequencing data. The files were converted to BedGraph format using the bigWigtoBedGraph tool[33]. Biological replicates were combined using BEDTools unionbedg[34], where only regions present in both replicates were retained, and their signal intensities averaged to provide a single score per genomic region. The datasets were intersected with the 160,276 genomic regions that matched those used in the raw count matrix of Gabriel et al.[11] using BEDTools intersect[34]. For each overlapping region, a weighted average accessibility score was calculated, considering both the signal intensity and the overlap length between the predefined genomic regions and the observed signal.

4.4 Pre-processing of cfDNA-Seq datasets

The cfDNA-Seq datasets were aligned on the human genome (hg38) and the fragment centers positions were calculated for the reads on the forward strand. Fragments with a length less than 120 bp or greater than 200 bp were excluded. The fragment center counts were summed at each genomic position. Cell-type specific marker regions including 716 regions defined by Gabriel et al.[11] and an expanded set of 1035 regions identified through differential analysis in this study were first trimmed to 2,000 bp by extending 1,000 bp in both directions from the midpoint. BEDTools map[34] was used to intersect the fragment center counts with these regions. To ensure comparability, z-score normalization was applied to the 2,000 bp bins followed by Whittaker smoothing (lambda: 1000, order: 2) and Gaussian smoothing (sigma: 30) to get a continuous fragment center score (Figure 5). The smoothed data was trimmed to 100 bp around the center of each region, and the mean fragment center counts within this region was calculated to focus on the most informative signal, as the central region is expected to capture the cell-type specific fragment distributions. To obtain a proxy of chromatin accessibility, the mean fragment center

counts were inverted and min-max scaled to a range of 0 to 1. The final accessibility scores were assigned back to the original marker regions for the downstream analysis.

4.5 Pairwise differential accessibility analysis

The pre-processed ATAC-Seq bulk datasets of hepatocytes, BRCA, COAD, LUAD, and LUSC were merged with the raw count matrix provided by Gabriel et al.[11] containing B cells, CD4+ T cells, CD8+ T cells, NK cells, macrophages, DCs, neutrophils, endothelial cells, and fibroblasts. To identify new cell-type specific markers, pairwise differential accessibility analysis was performed using the DESeq2 R package[35]. Normalization for sequencing depth was applied using the median ratio normalization (MRN) method implemented in DESeq2. The geometric mean of the counts across all samples was calculated for each region, and sample-wise size factor were determined as the median of the count ratios relative to this geometric mean. To compare the chromatin accessibility between each pair of cell types the Wald test was used as implemented in DESeq2. The resulting log2 fold change (log2FC) values expressed the magnitude of the differential accessibility, and the adjusted p-values corrected for multiple testing using the Benjamini-Hochberg method. To identify the new marker regions, all regions with log2fold change higher than 1.5 and adjusted p-values less than 0.1 were selected and ranked by their maximum adjusted p-values across all pairwise comparisons. Regions located on sex chromosomes were removed. The top 500 features (with the lowest adjusted p-values) were considered as candidate cell-type specific marker regions. Zhang et al.[12] identified modules of open chromatin regions accessible across multiple cell types or specific human tissues. Therefore, these regions were used to refine the set of marker regions based on their overlap. More precisely, regions with an overlap of more than 15% with immune (modules 8 to 25), endothelial (modules 26 to 35), and stromal (modules 41 to 49 and 139 to 150) were removed. Regions with an overlap of more than 60% with all other universal modules were filtered out. The new 319 cell-type specific marker regions for hepatocytes,

BRCA, COAD, LUAD, and LUSC were extracted and merged with the 716 original set of marker peaks provided by Gabriel et al. resulting in a new set of 1035 cell-type specific marker regions. To assess the specificity and potential batch effect issues, UMAP (Figure 2D) was run based on the 1035 cell-type specific regions after full quantile normalization (FQ-FQ) implemented in the EDASeq R package[36] to correct for depth and GC biases.

4.6 EPIC-ATAC deconvolution framework

Deconvolution is the process of estimating the cell-type proportions contributing to the DNA signal present in a given biological sample. The EPIC-ATAC deconvolution framework[11] was developed using ATAC-Seq bulk reference profiles and cell-type specific marker regions derived from purified samples of major immune, stromal, and vascular cell types that contribute to the tumor microenvironment. The code to perform the deconvolution using EPIC-ATAC is freely available as an R package[11]. Since cfDNA fragments are preferentially higher in closed chromatin regions while being lower in open chromatin regions, there is an inverse relationship between chromatin accessibility in the ATAC-Seq and cfDNA-Seq data distribution. This inverse relationship was used to integrate the cfDNA data into the EPIC-ATAC deconvolution framework, because cfDNA fragment center counts could potentially serve as a proxy for chromatin accessibility and resemble ATAC-Seq data when inverted. To infer the cell-type proportions, EPIC-ATAC uses a constrained least square optimization approach, where the observed signal is decomposed into weighted contributions of each reference cell type. The optimization is constrained to ensure that cell-type proportions remain non-negative and sum to one. In addition, it considers the variability within the cell-type specific marker regions to assign higher weights to the regions with lower variability.

4.7 Synthetic sample generation

Since cfDNA samples contain a mixture of cell types, the ground truth of the cell type composition is unknown. Therefore, synthetic samples with defined compositions were generated using the 9x716 cell type marker matrix of EPIC-ATAC. Predefined proportions of 0%, 1%, 2%, 4%, 8%, 13%, 17%, 21% and 34% were assigned to each cell type to systematically vary their contribution. To generate the synthetic datasets, the cell type marker matrix values were multiplied by the synthetic composition matrix to ensure that each sample generated reflected a controlled mixture of the cell types. This approach resulted in 72 unique combinations of cell types and proportions, providing a structured dataset for evaluating the deconvolution performance of EPIC-ATAC (Figure 3).

4.8 Down-sampling for assessing the deconvolution performance

Compared to the synthetic samples generated in this study, the cfDNA samples from Cristiano et al.[15] have significantly lower sequencing coverage (2x – 8x). To determine the impact of sequencing depth on the deconvolution performance, synthetic samples were down sampled to simulate different coverage levels (0.1x, 0.3x, 1x, 3x, 9x, 30x, 80x, and 245x) while maintaining the synthetic proportional cell type contribution. The synthetic matrix was transformed by dividing all values by the lowest nonzero value in the dataset. This approach ensured that the minimum nonzero value was set to 1, while the zero values remained zero allowing for proportional scaling. To determine the appropriate coverage levels, the total summed signal of the scaled synthetic matrix was calculated. To derive the scaling factors, the total summed signal of the scaled synthetic matrix was compared to the average total cfDNA fragment center counts across five healthy cfDNA samples. By proportionally adjusting this average cfDNA signal based on the desired sequencing depth, the equivalent signal at the target coverage was calculated. In other words, relative to the baseline 2.7x coverage of the healthy cfDNA samples, the equivalent signal represents the expected total number of cfDNA fragment

center counts that would be observed at a given sequencing depth. The ratio of the total summed signal of the scaled synthetic matrix to the equivalent signal at each target coverage was used to calculate the scaling factors: 0.1x (2827.08), 0.3x (942.36), 1x (282.71), 3x (94.24), 9x (31.41), 30x (9.42), 80x (3.53), 245x (1.15). These scaling factors ensured that the down-sampled synthetic datasets accurately reflected the different coverage levels allowing for a systematic evaluation of how reduced coverage affects the cfDNA-based deconvolution with EPIC-ATAC (Figure 4).

4.9 Building the new reference profiles

To build the reference profile for the EPIC-ATAC framework with the extended cell types, transcript per million (TPM) like normalization was applied. It has been shown that TPM transformation was suitable for the deconvolution of cell fractions from RNA-Seq bulk mixtures[37,38]. The raw read counts were divided by the region length, corrected for sequencing depth, and rescaled to the total number of counts per sample added up to 10^6. For each region, the median of the TPM-like transformed counts across all samples of a cell type was calculated to form the reference profile compatible with EPIC-ATAC. To account for variability in chromatin accessibility across samples, the interquartile range (IQR) of the TPM-like transformed counts for each region in each cell type was calculated. This measure has been integrated into the constrained least squares optimization of EPIC-ATAC to improve the robustness of the deconvolution[37]. To complete all requirements for EPIC-ATAC, the set of significant peaks as well as the list of the extended cell-type specific marker regions were integrated into the framework.

5 Supporting Information

5.1 List of abbreviations

ATAC-Seq	Assay for Transposase-Accessible Chromatin using sequencing
BRCA	Breast cancer
CCC	Concordance Correlation Coefficient
cfDNA	cell-free DNA
COAD	Colon Adenocarcinoma
ctDNA	Circulating tumor DNA
DCs	Dendritic Cells
EPIC	Estimating the Proportions of Immune and Cancer cells
LUAD	Lung Adenocarcinoma
LUSC	Lung Squamous Cell Carcinoma
NK	Natural Killer
RMSE	Root Mean Square Error
TPM	Transcript Per Million
UMAP	Uniform Manifold Approximation and Projection

5.2 Code availability

The code to pre-process publicly available ATAC-Seq and cfDNA-Seq samples, perform deconvolution with EPIC-ATAC, as well as the code used to identify the new cell-type specific marker regions and build the reference profile is available on GitHub: (https://github.com/DDE277/master-thesis-project).

References

1. Crosby, D. *et al.* Early detection of cancer. *Science* **375**, eaay9040 (2022).
2. Phallen, J. *et al.* Direct detection of early-stage cancers using circulating tumor DNA. *Science Translational Medicine* **9**, eaan2415 (2017).
3. Hawkes, N. Cancer survival data emphasise importance of early diagnosis. *BMJ* **364**, l408 (2019).
4. Wan, J. C. M. *et al.* Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat Rev Cancer* **17**, 223–238 (2017).
5. Stewart, C. M. & Tsui, D. W. Y. Circulating cell-free DNA for non-invasive cancer management. *Cancer Genet* **228–229**, 169–179 (2018).
6. Sender, R., Noor, E., Milo, R. & Dor, Y. What fraction of cellular DNA turnover becomes cfDNA? *eLife* **12**, RP89321 (2024).
7. Penny, L., Main, S. C., De Michino, S. D. & Bratman, S. V. Chromatin- and nucleosome-associated features in liquid biopsy: implications for cancer biomarker discovery. *Biochem Cell Biol* **102**, 291–298 (2024).
8. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* **164**, 57–68 (2016).
9. Moss, J. *et al.* Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat Commun* **9**, 5068 (2018).
10. Li, S. *et al.* Comprehensive tissue deconvolution of cell-free DNA by deep learning for disease diagnosis and monitoring. *Proceedings of the National Academy of Sciences* **120**, e2305236120 (2023).

11. Gabriel, A. A., Racle, J., Falquet, M., Jandus, C. & Gfeller, D. Robust estimation of cancer and immune cell-type proportions from bulk tumor ATAC-Seq data. *eLife* **13**, (2024).

12. Zhang, K. *et al.* A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**, 5985 (2021).

13. Fox-Fisher, I. *et al.* Remote immune processes revealed by immune-derived circulating cell-free DNA. *eLife* **10**, e70520 (2021).

14. Ma, L. *et al.* Liquid biopsy in cancer: current status, challenges and future prospects. *Sig Transduct Target Ther* **9**, 1–36 (2024).

15. Cristiano, S. *et al.* Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **570**, 385–389 (2019).

16. Fox-Fisher, I. *et al.* Remote immune processes revealed by immune-derived circulating cell-free DNA. *eLife* **10**, e70520 (2021).

17. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet* **48**, 1193–1203 (2016).

18. Trizzino, M. *et al.* EGR1 is a gatekeeper of inflammatory enhancers in human macrophages. *Science Advances* **7**, eaaz8836 (2021).

19. Leylek, R. *et al.* Chromatin Landscape Underpinning Human Dendritic Cell Heterogeneity. *Cell Reports* **32**, 108180 (2020).

20. Ram-Mohan, N. *et al.* Profiling chromatin accessibility responses in human neutrophils with sensitive pathogen detection. *Life Science Alliance* **4**, (2021).

21. Xin, J. *et al.* Chromatin accessibility landscape and regulatory network of high-altitude hypoxia adaptation. *Nat Commun* **11**, 4928 (2020).

22. Ge, X. *et al.* Functional genomics atlas of synovial fibroblasts defining rheumatoid arthritis heritability. *Genome Biology* **22**, 247 (2021).

23. Ma, H. *et al.* The nuclear receptor THRB facilitates differentiation of human PSCs into more mature hepatocytes. *Cell Stem Cell* **29**, 1611 (2022).

24. Collins, J. M. *et al.* Regulatory variants in a novel distal enhancer regulate the expression of CYP3A4 and CYP3A5. *Clin Transl Sci* **15**, 2720–2731 (2022).

25. Wu, H. *et al.* Integrative omics analysis reveals gene regulatory mechanisms distinguishing organoid-derived hepatocytes from primary human hepatocytes. 2023.12.05.570132 Preprint at https://doi.org/10.1101/2023.12.05.570132 (2023).

26. An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* **489**, 57–74 (2012).

27. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**, eaav1898 (2018).

28. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

29. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

30. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).

31. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).

32. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**, 9354 (2019).

33. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res* **12**, 996–1006 (2002).

34. BEDTools: a flexible suite of utilities for comparing genomic features | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/26/6/841/244688.

35. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).

36. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-Content Normalization for RNA-Seq Data. *BMC Bioinformatics* **12**, 480 (2011).

37. Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E. & Gfeller, D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife* **6**, e26476 (2017).

38. Sturm, G. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35**, i436–i445 (2019).