

High-resolution reconstruction of cell-type-specific transcriptional regulatory processes from bulk sequencing samples

Received: 16 April 2025

Accepted: 11 June 2026

Published online: 13 July 2026

 Check for updates

Li Yao (姚力) ^{1,2}, Sagar R. Shah ^{2,3}, Abdullah Ozer ³, Yutong Zhu ^{1,2},
Xiuqi Pan^{2,3}, Tianyu Xia^{1,2}, Junke Zhang ^{1,2}, Alden King-Yung Leung^{1,2},
Meihan Wei^{1,2}, John T. Lis^{1,3} & Haiyuan Yu (于海源) ^{1,2} ✉

Single-cell sequencing methods such as scRNA-seq and scATAC-seq have advanced our understanding of individual cellular functions but experimentally adapting genome-wide assays measuring other genomic features to achieve single-cell resolution remains a technical challenge. Here we introduce deep-learning-based deconvolution of tissue profiles with accurate interpretation of locus-specific signals (DeepDETAILS), a quas-supervised framework performing cross-modality deconvolution using scATAC-seq reference libraries for other bulk datasets. DeepDETAILS enables base-pair-resolution mapping of genomic signals across diverse cell types, with great versatility for various omics datasets, including nascent transcript sequencing (such as PRO-cap and PRO-seq) and ChIP-seq for chromatin modifications. Using DeepDETAILS, we generated a compendium of high-resolution nascent transcription and histone modification signals across 39 diverse human tissues and 86 distinct cell types. Furthermore, we applied our compendium to fine-map risk variants associated with primary sclerosing cholangitis, a progressive cholestatic liver disorder, and revealed a potential etiology of the disease.

Transcription is a tightly regulated process involving multiple rate-limiting steps, including transcription activator binding, chromatin remodeling and modification, enhancer–promoter (E–P) interactions, preinitiation complex formation, promoter escape, promoter-proximal pausing and release of paused Pol II into productive elongation^{1,2}. Understanding the regulatory mechanisms governing transcription is crucial for deciphering complex biological processes. High-throughput sequencing assays have enhanced our understanding by profiling transcription factor (TF) binding, histone modifications³, E–P interactions⁴ and nascent transcription dynamics^{5–7}. Interplay among regulatory layers⁸ enables assays designed for profiling one process to capture signals from others. For example, GRO/PRO-cap^{5,9} were developed to map all active transcription start sites (TSSs)

genome-wide but the discovery of enhancer transcription¹⁰ has made them highly sensitive assays for pinpointing active enhancers^{11,12}. Collectively, these bulk sequencing technologies provide tools to study the transcriptional regulatory processes that control cellular function and fate.

While powerful, bulk sequencing assays aggregate signals across cell populations, which can obscure cell-type-specific regulatory mechanisms when applied to tissue samples usually composed of diverse cell types that are available in limited quantities. In contrast, scRNA-seq, snRNA-seq, scATAC-seq and snATAC-seq^{13,14} resolve cellular heterogeneity and enable the characterization of regulatory processes at the individual cell level. In addition to measuring mRNA expression and chromatin accessibility, recent advances have enabled the detection of

¹Department of Computational Biology, Cornell University, Ithaca, NY, USA. ²Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY, USA. ³Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA. ✉e-mail: haiyuan.yu@cornell.edu

more genomic features^{15–19}. However, adapting bulk assays to single-cell resolution in general faces substantial experimental challenges, including the scarcity of input material and limitations in the efficiency of library preparations that can hinder certain assay functionalities¹⁹. Moreover, ease of use^{16,17} and cost²⁰ present additional obstacles.

Alongside the development of single-cell-resolution assays, there is increasing interest in developing *in silico* deconvolution methods to reconstruct cell-type-specific profiles from bulk sequencing. Current methods primarily rely on single-cell data from the same assay type as references to deconvolute bulk data from the same modality. A common approach uses scRNA-seq reference count matrices to estimate cell type fractions^{21,22} and cell-type-specific gene expression from bulk RNA-seq samples^{23–27}. Because these methods require panels of bulk or reference samples, they have mainly been applied to consortium-level data such as TCGA^{23,24} and SFARI^{26,28}. Recently works have expanded to deconvolve other modalities, including DNA methylation^{29,30}, using matched-modality references. However, obtaining single-cell sequencing reference libraries for many other omics modalities (for example, nascent transcriptome) remains costly and challenging. Given the interconnected nature of regulatory layers, an alternative approach is to use single-cell references from more widely available modalities, such as scATAC-seq and snATAC-seq, to deconvolve bulk data from assays that are not single cell ready (that is, cross-modality deconvolution). However, a major challenge in cross-modality deconvolution is the lack of direct correspondence or alignment between different omics measurements, which can lead to discrepancies in data interpretation. For instance, a genomic region may exhibit similar accessibility across multiple cell types but only initiate transcription in one.

In this study, we developed DeepDETAILS, a deep learning model enabling precise deconvolution of bulk omics profiles into cell-type-specific signals at base-pair resolution in a cross-modality manner. DeepDETAILS deconvolves bulk sequencing data—such as nascent transcriptome and histone modification profiles—using scATAC-seq or snATAC-seq libraries from the same type of tissue as reference. Our results demonstrate that DeepDETAILS outperformed traditional statistical deconvolution methods. Using DeepDETAILS, we developed a compendium of high-resolution nascent transcription and histone modification maps across 39 human tissues and 86 distinct cell types. Furthermore, we applied our compendium to fine-map risk variants associated with primary sclerosing cholangitis, a challenging illness with unknown etiology, characterized by chronic bile duct destruction and progression to end-stage liver disease, revealing potential cellular origins of the disease.

Results

Deep learning architecture of DeepDETAILS

As fundamental building blocks of eukaryotic genome architecture, the placement of nucleosomes influences many transcriptional regulatory processes. Genome-wide landscapes of nucleosome-depleted regions or open chromatin regions indicate active regulatory regions that are highly cell type specific^{31–33}. With scATAC-seq and snATAC-seq becoming increasingly available, we sought to use chromatin accessibility information for each cell type to reconstruct other regulatory processes from bulk sequencing assays. To achieve this, we developed DeepDETAILS, a deep learning framework designed to reconstruct cell-type-specific regulatory signals from bulk data (Fig. 1a, Methods and Supplementary Notes). DeepDETAILS uses branches of dilated convolutional neural network (CNN) blocks to make individual predictions for each cell type, which are then combined linearly to reconstruct the bulk signal. Although deep learning is widely used in genomics, most applications are fully supervised, learning regulatory sequence patterns³⁴ or predicting noncoding variant effects^{35,36} from experimentally measured target signals, such as transcription initiation³⁷ or histone modifications³⁸ (Extended Data Fig. 1a). In contrast, DeepDETAILS is designed for deconvolution, where cell-type-specific signals are not

directly available. Instead of learning from ground-truth signals for each cell type, DeepDETAILS operates in a quasisupervised manner by minimizing the differences between predicted and observed bulk signals (Extended Data Fig. 1a).

Inspired by the fact that open chromatin is necessary but not sufficient for many regulatory processes (Extended Data Fig. 1b–g), we introduced a constraint based on local accessibility: if a locus is not accessible in one cell type, tentative predictions made for this cell type will not be incorporated into the final prediction. This acts as a feedback mechanism during training and introduces asymmetry into the architecture.

We implemented two versions of DeepDETAILS (Fig. 1a, Extended Data Fig. 2a–e and Methods). The ‘seq-only’ model uses only DNA sequence and chromatin accessibility constraints to maximize the interpretability of sequence contributions. The ‘fused’ model combines base-pair-resolution chromatin accessibility profiles learned by gated recurrent units (GRUs) with sequence embeddings learned by CNNs. Incorporating accessibility embeddings enabled a smaller sequence-processing architecture, reducing GPU memory usage (Extended Data Fig. 2f) and improving scalability for deconvolving more cell types under the same hardware configuration (Extended Data Fig. 2g).

DeepDETAILS learns TF motifs that govern cell-type-agnostic general regulatory processes and cell identity

To validate DeepDETAILS, we conducted a proof-of-concept study using a simulated heterogeneous ‘tissue’ sample composed of K562, GM12878 and MCF-7 (Extended Data Fig. 2h). We generated the bulk library by combining GRO/PRO-cap data, which captures transcription initiation signals from all enhancers and promoters genome-wide, for each cell line, and used snATAC-seq of these cell lines as the reference (Extended Data Fig. 2h and Methods). Both the seq-only and the fused versions of DeepDETAILS successfully reconstructed the initiation profiles for all three cell types (Fig. 1b and c).

To interpret the sequence motifs that DeepDETAILS used in its predictions, we used the input $\times$ gradient method³⁹. The framework identified motifs associated with transcription initiation in general, including binding sites for Sp/KLF and ETS family of TFs (Fig. 1b), as well as initiator and TATA box (Extended Data Fig. 3a), consistent with previous studies^{37,40,41}. Additionally, DeepDETAILS captured cell-type-specific motifs reflective of cell identity. For instance, GATA1 is a pioneer TF critical for erythroid differentiation and is notably active in K562 cells⁴². DeepDETAILS learned prominent GATA1-like motifs in K562 cells but not in other cell types (Fig. 1c). IRF-family TFs are known to have critical roles in lymphocyte development⁴³ and DeepDETAILS also identified IRF-like motif contribution in GM12878 (Extended Data Fig. 3b). While both models produced similar deconvolution results, the seq-only model demonstrated better learning of DNA syntax than the fused model, likely because the fused model can partially use accessibility embeddings to infer cell type identity and was designed with a smaller architecture for efficiency. Given its hardware-friendly design, we proceeded with the fused version for subsequent analyses.

DeepDETAILS can deconvolve regulatory signals from a wide variety of bulk sequencing assays

We assessed DeepDETAILS across distinct bulk sequencing assays, including PRO-cap for transcription initiation (Figs. 1b,c and 2a), ChIP-seq for H3K27ac, H3K4me1 and H3K4me3 histone modifications (Fig. 2b–d and Extended Data Fig. 4a–c) and PRO-seq for pause–release dynamics (Extended Data Fig. 4d). Each assay captures different aspects of regulatory processes at varying resolutions; ChIP-seq provides broader signals, while PRO-cap/seq offers base-pair resolution. Despite these differences in assay types and data characteristics, DeepDETAILS consistently performed well across all assays, demonstrating its robustness and versatility.

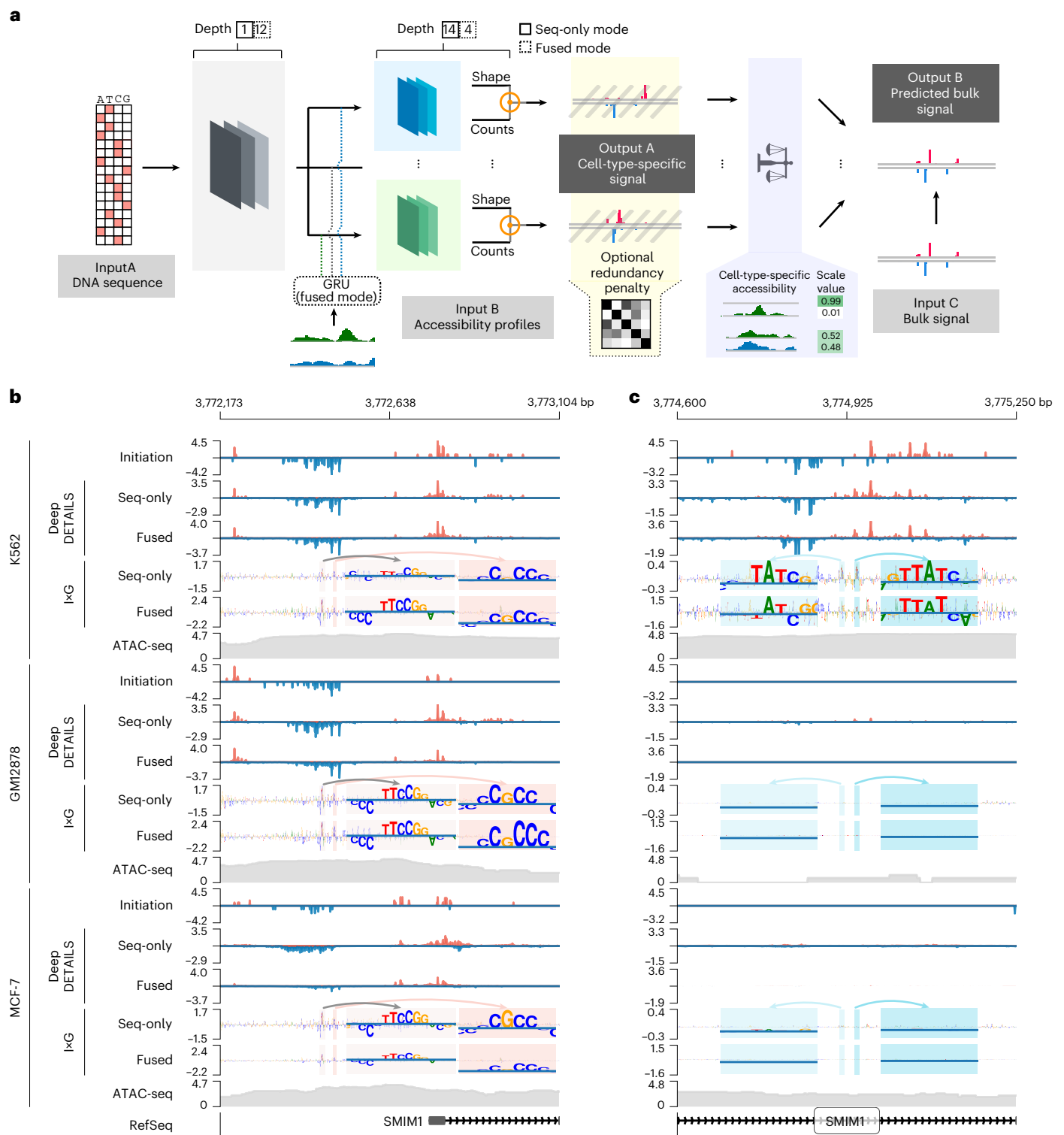


Fig. 1 | Architecture of DeepDETAILS. a, Schematic overview of the DeepDETAILS framework. The model takes three inputs: (1) DNA sequence information; (2) cell-type-specific accessibility profiles derived from scATAC-seq and snATAC-seq data; and (3) 'bulk' sequencing signals. DeepDETAILS is optimized to minimize the difference between real and predicted bulk tracks. Components unique to the fused mode of DeepDETAILS are indicated by dashed boxes. **b, c**, DeepDETAILS deconvolved a simulated bulk PRO-cap library consisting of three cell types: K562, GM12878 and MCF-7. The ground-truth signals are labeled as 'Initiation', while deconvolved signals from the seq-only and fused models are shown separately. Sequence motifs contributing to the

prediction in each cell type were shown in the '1 × G' (input × gradient) tracks. **b**, Motifs known to be important for transcription initiation in general are highlighted with pink shading. **c**, Motifs associated with cell-type-specific identity are highlighted with blue shading (for example, GATA1-like motif for K562, with the consensus sequence GATAA/TTATC). Unless explicitly specified, inverse hyperbolic sine (asinh) transformation was applied to initiation and chromatin accessibility by default to account for their large dynamic range (for example, between promoters versus enhancers). The asinh transformation is similar to log transformation but supports negative values.

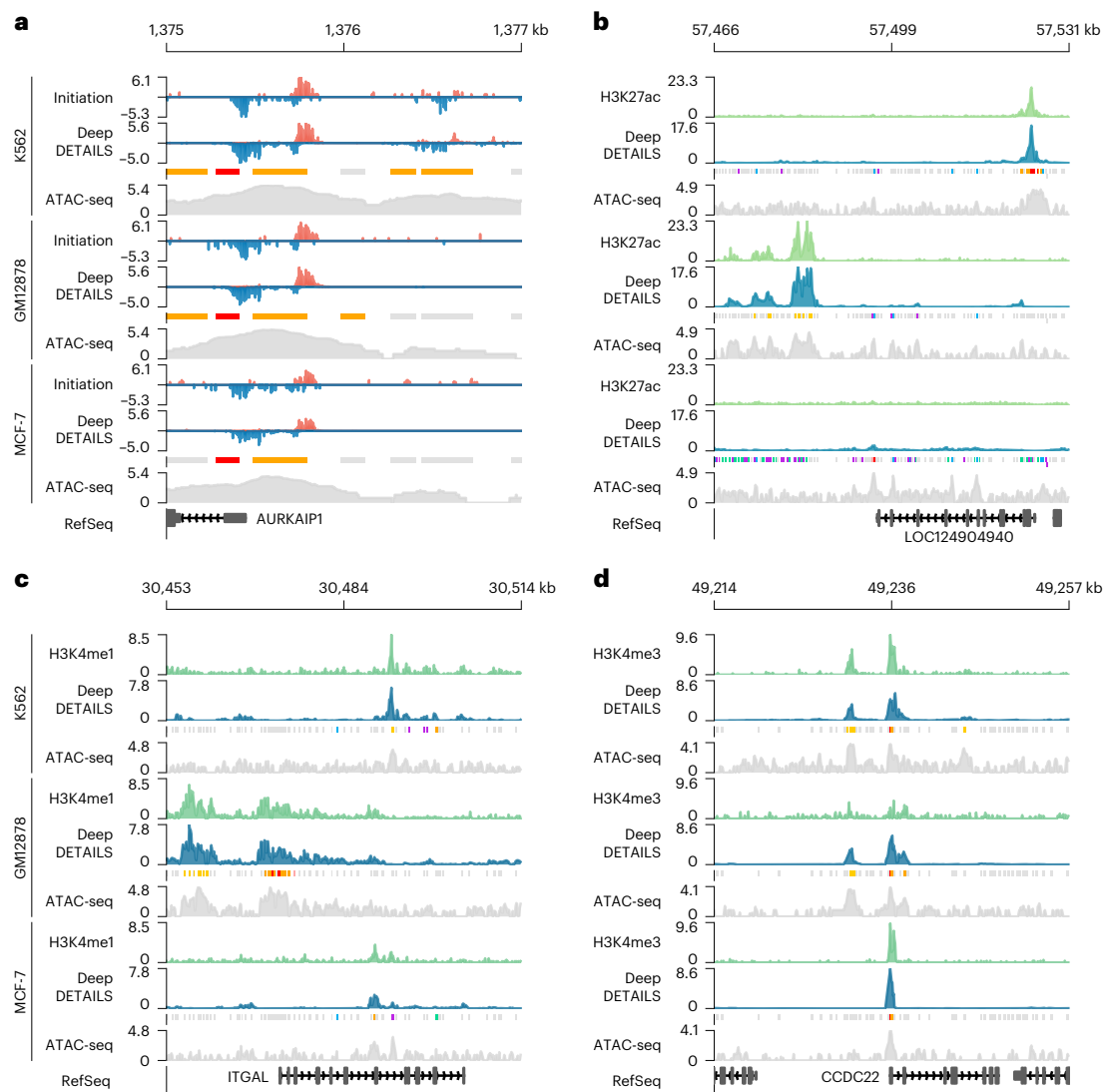


Fig. 2 | DeepDETAILS works for multiple bulk sequencing assays.

a, DeepDETAILS deconvolved simulated bulk transcription initiation profiles captured by GRO/PRO-cap from both promoters (labeled in red boxes) and enhancers (labeled in orange boxes). **b–d**, DeepDETAILS also successfully

deconvolved bulk histone modification signals, including for H3K27ac (**b**), H3K4me1 (**c**) and H3K4me3 (**d**). Candidate regulatory elements such as enhancers and promoters were annotated on the basis of ENCODE cCRE definitions⁷².

Deconvolution by DeepDETAILS matches or exceeds performance of canonical supervised direct prediction methods

DeepDETAILS operates in a deconvolution setting, where it only has access to bulk signals and not the desired cell-type-specific prediction targets. Despite facing a more challenging learning task than previous deep learning studies (Extended Data Fig. 1a), we evaluated its performance against published fully supervised models for sequence learning^{37,38}. We created simulated bulk sequencing libraries by combining signals from five different cell lines (Methods) chosen to reflect higher or lower pairwise similarity in chromatin accessibility profiles when such data were available (Extended Data Fig. 5a). This approach aimed to assess how effectively DeepDETAILS handles samples with varying cell population structures. For example, we simulated two sets of bulk libraries capturing nascent transcription initiation: A673, HCT116, HUVEC, K562 and MCF 10A (more similar group) and Caco-2, Calu3, GM12878, LNCaP and MCF-7 (more distinct group) (all simulation formulas available in Supplementary Table 2). We used DeepDETAILS to reconstruct signal tracks of initiation, pause–release and histone H3K4me3 or H3K27ac modifications for each cell line and compared its performance to two state-of-the-art supervised models: Puffin-D³⁷

and Enformer³⁸. Puffin-D was developed for predicting transcription initiation, which we also repurposed to study pause–release by training separate models using the ground truth (Methods). Predicted profiles of histone modifications from Enformer were retrieved from the original study³⁸. As Enformer predicts signals at 128-bp resolution, we binned the predictions from both DeepDETAILS and Puffin-D to match the resolution. We compared predictions to actual signal tracks for each cell type using Pearson's r , as reported in the original studies (Fig. 3a). DeepDETAILS achieved similar performances to state-of-the-art supervised models (Pearson's r for DeepDETAILS: 0.704 (initiation), 0.730 (pause–release) and 0.767 (histone modification); other models: 0.726, 0.734 and 0.679, respectively) (Fig. 3a). Because DeepDETAILS and Puffin-D make predictions at base-pair resolution and PRO-cap/3' of CoPRO yields base-pair-resolution profiles, we tested performances recovering the dominant TSSs and pausing sites at 1-bp resolution. We demonstrated that DeepDETAILS achieved comparable performance to state-of-the-art models with large receptive fields in predicting dominant TSSs and pausing sites. Specifically, DeepDETAILS obtained Pearson's r values of 0.800 for TSSs and 0.837 for pausing sites, closely matching the 0.817 and 0.836 scores of Puffin-D (Fig. 3b,c

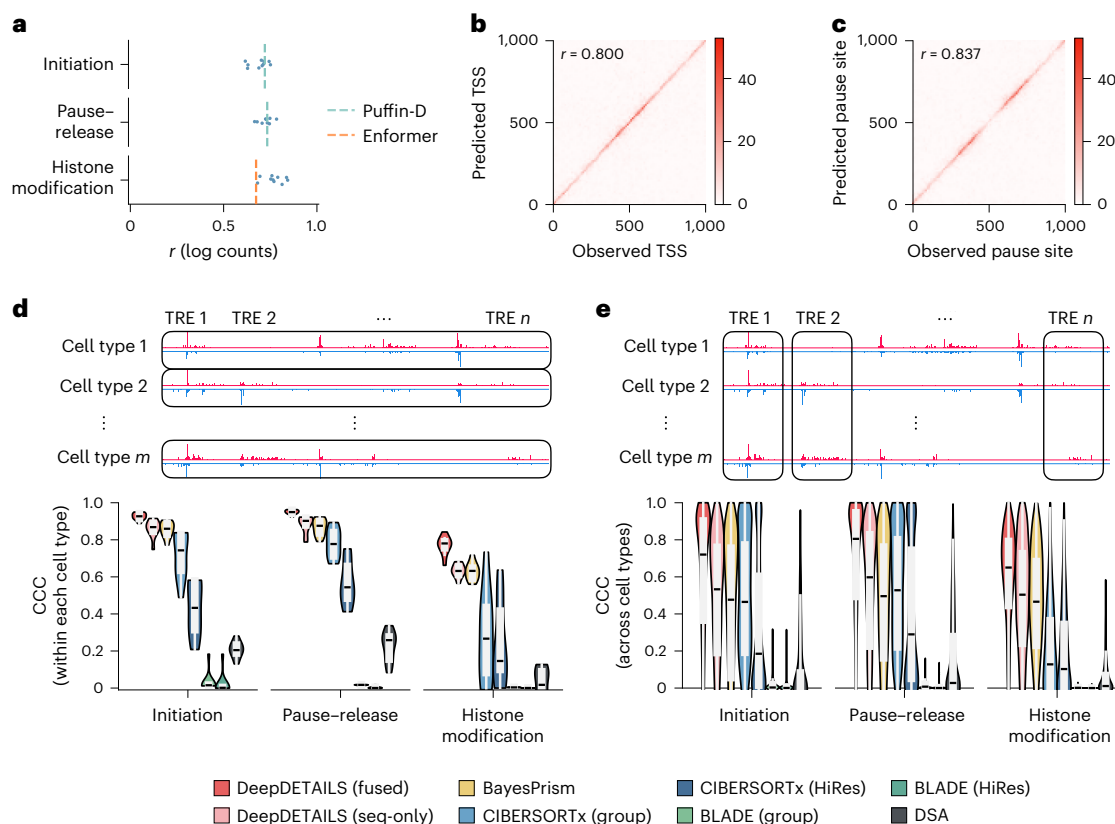


Fig. 3 | Performance comparison of DeepDETAILS to state-of-the-art methods.

a, DeepDETAILS was compared to state-of-the-art supervised deep learning models, Puffin-D (for transcription initiation/pause–release) and Enformer (for histone modification). The evaluation used Pearson’s r at 128-bp resolution with log-transformed read counts. Dashed line shows the median of r that the corresponding supervised model archived ($n = 10$). **b**, Pearson’s r was calculated to evaluate consistency between the DeepDETAILS reconstructed dominant TSSs and actual dominant TSSs in the ground-truth GRO/PRO-cap libraries. The analysis sampled 5,000 peak regions for this assessment. **c**, Similarly, Pearson’s r was used to assess consistency between the DeepDETAILS

reconstructed dominant pause sites and actual dominant sites at the 3’ end of the CoPRO libraries. This evaluation also used a sample of 5,000 peak regions. **d,e**, Performance comparison between DeepDETAILS and statistical deconvolution tools (1-kb resolution). Black boxes indicate type of CCC evaluation. In the box plots, the center lines, box limits and whiskers denote the median, upper and lower quartiles and $1.5 \times$ the interquartile range (IQR), respectively. **d**, Within-each-cell-type evaluation ($n = 10$ for each bar). **e**, Across-cell-type evaluation, with sample sizes of $n = 169,861$ for initiation, $n = 165,920$ for pause–release and $n = 163,888$ for histone modification for each bar. TRE, transcriptional regulatory element.

and Extended Data Fig. 5b,c). This performance is notable given that DeepDETAILS was trained on simulated bulk data (mixture of five cell lines), whereas Puffin-D and Enformer were trained on ground-truth bulk data for each cell line.

DeepDETAILS outperforms statistical deconvolution tools in cross-modality settings

Statistical tools are widely used in deconvolving bulk profiles with references of the same modality^{23–27}. These tools consider bulk sequencing profiles as linear combinations of signals from each cell type and calculate cell-type-specific signals by solving a system of equations^{23–27}. However, references may not be available for assays that are technically difficult to perform at single-cell resolution (for example, PRO-cap). DeepDETAILS addresses this limitation by enabling cross-modality deconvolution. It uses references from one modality (scATAC-seq and snATAC-seq) to deconvolve bulk profiles of another modality (for example, PRO-cap), thereby overcoming the hurdle of unavailable references for certain assays.

To evaluate performance in cross-modality deconvolution, we compared DeepDETAILS to several statistical deconvolution methods (BayesPrism²⁴, CIBERSORTx²³, BLADE²⁵ and DSA⁴⁴) using simulated bulk samples (Fig. 3d,e). Although DeepDETAILS provides base-pair-resolution predictions, statistical tools typically generate count tables for genes or regions. For a fair comparison, we summed

DeepDETAILS predictions over 1-kilobase (kb) regions and used statistical tools to reconstruct read counts tables in the same regions (Methods). DeepDETAILS demonstrated superior performance (Fig. 3d,e), achieving the highest Lin’s concordance coefficients (CCCs) in within-cell-type comparisons: 0.927 for initiation, 0.950 for pause–release and 0.780 for histone modifications. The best-performing statistical methods yielded CCCs of 0.860, 0.876 and 0.632, respectively (Fig. 3d). DeepDETAILS also excelled in across-cell-type comparisons, with CCCs of 0.717, 0.803 and 0.651 for initiation, pause–release and histone modifications, outperforming the statistical methods’ results of 0.477, 0.530 and 0.466 (Fig. 3e and Extended Data Fig. 5d). We also assessed prediction error using root-mean-squared error (r.m.s.e.); DeepDETAILS reconstructed read counts with the highest accuracy, with a mean overall r.m.s.e. of 0.944 (Extended Data Fig. 5e). Additionally, statistical deconvolution methods usually require multiple bulk samples^{23–25,44} or reference samples²⁶, DeepDETAILS only needs one bulk and reference sample, which enables its broader applications in research projects of different scales and purposes.

DeepDETAILS works robustly in complex settings

Many factors can impact the accuracy and reliability of deconvolution outcomes. Therefore, we aimed to evaluate the performance of DeepDETAILS under more demanding conditions. First, we examined the robustness of DeepDETAILS when applied to libraries with varying

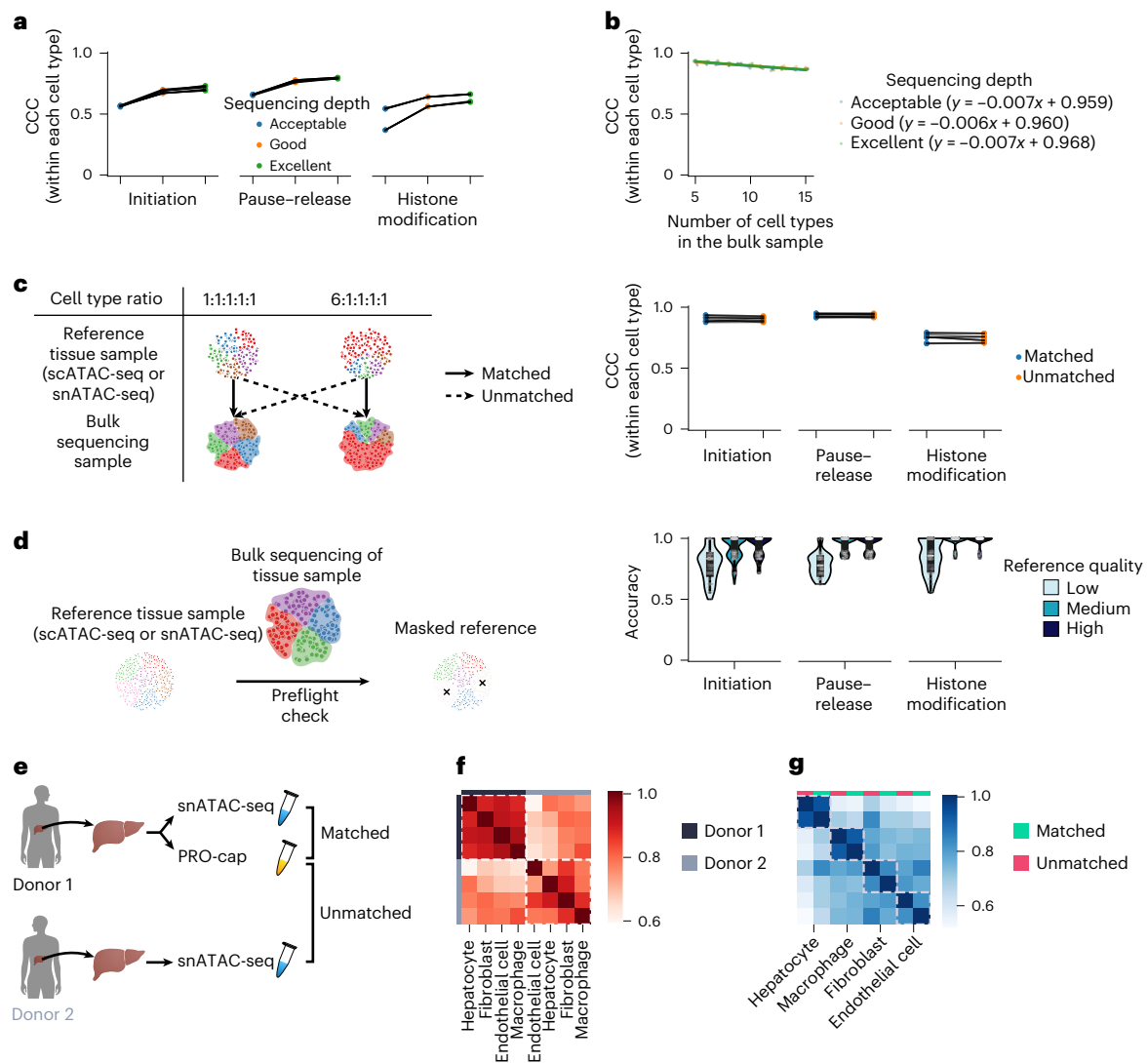


Fig. 4 | Performance evaluation of DeepDETAILS under challenging cases.

a, DeepDETAILS performance by varying the sequencing depths of bulk and reference libraries. For each sequencing technique, two sets of bulk–reference pairs were simulated and deconvolution was repeated three times ($n = 6$ per technique). **b**, Evaluation of DeepDETAILS performance on deconvolving bulk transcription initiation sequencing libraries with different numbers of cell types in the bulk and reference ($n = 3$ for each simulated setting). **c**, Schematic (left) and results (right) of using DeepDETAILS to deconvolve bulk samples with unmatched reference libraries (ratios of cell types do not match in the bulk and reference; $n = 12$). For **a–c**, evaluation was performed by calculating the CCC

within each cell type as in Fig. 3d. **d**, Schematic of the preflight check step (left) and the accuracy of preflight in masking out cell types in the reference but not in the bulk (right; $n = 126$ for each condition). In the box plot, the center lines, box limits and whiskers denote the median, upper and lower quartiles and $1.5 \times$ the IQR, respectively. **e**, Schematic of evaluating DeepDETAILS with real datasets. **f**, Genome-wide correlation analysis (Pearson's r) suggested the pairwise similarity relationships between different cell types were disrupted in the two reference samples. **g**, The deconvolved signals from DeepDETAILS maintained the pairwise similarity (Pearson's r) relationships regardless of the choice of reference libraries.

sequencing depths. We simulated bulk and reference libraries across three distinct sequencing depth categories: excellent, good and acceptable (Methods). DeepDETAILS maintained consistent performance levels, with minimal reductions observed as the sequencing depths of the libraries decreased. Specifically, using CCC as a metric, within-cell-type comparisons demonstrated limited performance decline when transitioning from excellent sequencing quality to good quality (range of decreasing: 0.002 to 0.042; Fig. 4a) and from good to acceptable quality (range of decreasing: 0.031 to 0.099, Fig. 4a). Similar trends were observed for across-cell-type comparisons (Extended Data Fig. 6a). These results indicate that DeepDETAILS exhibits robust performance across different sequencing depths, making it a reliable tool for deconvolution analysis even under suboptimal conditions.

Cell type composition in human tissues can be highly complex; therefore, a key attribute of deconvolution methods is their capacity to

effectively accommodate this variation in heterogeneity. We simulated bulk libraries containing 5–15 cell types and assessed DeepDETAILS performance across three sequencing depth conditions. We found that the number of cell types in the bulk library has only a minimal impact on both within-cell-type and across-cell-type CCCs. Specifically, the inclusion of one additional cell type resulted in a decrease of 0.007 in within-cell-type CCC (Fig. 4b) and 0.008 in across-cell-type CCC (Extended Data Fig. 6b). Under more imbalanced cellular compositions (dominant cells: ~90%, rare populations: 2.5–10%), DeepDETAILS maintained consistent across-cell-type concordance despite marginal within-cell-type CCC decreases ($\Delta = -0.002$ per 1% dominance increase; Extended Data Fig. 6c,d), preserving its core capability to accurately deconvolve signals. We also performed simulations by formulating bulk samples mimicking the similarity relationships between cell types in real-world normal and disease samples (Extended Data Fig. 6e,f and

Supplementary Notes) and we observed consistent and robust performance from DeepDETAILS (Extended Data Fig. 6g,h).

Previous simulations assumed that the bulk and reference samples originated from the same source and shared identical cell type ratios. Expanding the ability of DeepDETAILS to work with unmatched samples (that is, those from different sources or with varying cell type proportions) would enhance its practicality. To evaluate the capability of DeepDETAILS, we simulated two distinct groups of samples: one group featuring an even distribution of cell types and another dominated by a single cell type (comprising 60% of the total). ‘Matched deconvolution’ was defined as using reference and bulk samples from the same group, while ‘unmatched deconvolution’ involved pairing samples from different groups (Fig. 4c). The simulated unmatched settings were designed to represent scenarios with large fluctuations in cell type ratios. Our results showed consistent deconvolution outcomes in matched and unmatched settings, with differences in within-cell-type and across-cell-type CCCs ranging from 0.001 to 0.028 and 0.001 to 0.064, respectively (Fig. 4c and Extended Data Fig. 6i).

In addition to deconvolving cell type identity, we tested whether DeepDETAILS can dissect biological conditions beyond cell type identity. To do so, we obtained HCT116 cells with an auxin-inducible CTCF degradation system (LOC4335696 knock-in), performed PRO-cap under baseline (0 h) and perturbed (6 h auxin) conditions and generated simulated bulk mixtures containing both states together with three additional cell lines. We repeated this analysis across five distinct cell type compositions. Across these simulations, DeepDETAILS preserved state-specific information, with an average of 70.2% of TREs differentially expressed between the perturbed and unperturbed HCT116 states reconstructed in the correct direction (Extended Data Fig. 7a,b), demonstrating the capacity of DeepDETAILS to resolve biological context beyond static cell identity. In addition, we performed further simulations to evaluate the performance of DeepDETAILS with reference and bulk samples that exhibit varying biological conditions and library quality. We also assessed its robustness against biological variation. Our findings indicate that DeepDETAILS is resilient to out-of-distribution scenarios (Extended Data Fig. 7c–e and Supplementary Notes).

Consortium-level efforts^{3,45} have been increasing the availability of human tissue scATAC-seq and snATAC-seq libraries; thus, users may not need to generate their own reference libraries to deconvolve their bulk samples with DeepDETAILS. Reference scATAC-seq and snATAC-seq libraries provided by consortia are usually generated by pooling multiple specimens to ensure a comprehensive representation of different cell types; hence, there may be more cell types in the reference libraries than the bulk sample. To address this, DeepDETAILS uses a ‘preflight check’ step that estimates the proportions of different cell types within the bulk sample. It further identifies and masks out cell types present in the reference but absent in the bulk, preventing their influence on subsequent analyses (Methods). We evaluated the accuracy of the preflight step by blending in 1–5 additional cell types into the references used by DeepDETAILS. The results demonstrated high accuracy (median accuracy of 1.0) in detecting extraneous cell types when the reference libraries had adequate sequencing depth (Fig. 4d and Extended Data Fig. 7f).

Lastly, to further validate the robustness of DeepDETAILS for real-world applications, we applied this method to reconstruct cell-type-specific initiation maps using a bulk liver PRO-cap library (Fig. 4e). We used a matched snATAC-seq library from the same donor’s right lobe of the liver (ENCSTR364FHN). To further assess its performance in unmatched conditions, we deconvolved the same bulk library using another snATAC-seq dataset derived from a different donor’s liver (Fig. 4e). We observed substantial batch effects between the two snATAC-seq libraries, with accessibility tracks clustering by their sources rather than cell type identities (Fig. 4f). Additionally, there were noticeable disruptions in pairwise similarities between cell types

across donors. For instance, hepatocytes and endothelial cells showed high similarity (Pearson’s $r = 0.911$) in one donor, whereas fibroblasts displayed strong similarity to hepatocytes (Pearson’s $r = 0.896$) in another (Fig. 4f). Despite these challenges, DeepDETAILS demonstrated resilience by reconstructing consistent tracks across both matched and unmatched settings. The method achieved strong correlations for all cell types (Fig. 4g): hepatocytes (Pearson’s $r = 0.937$), macrophages ($r = 0.921$), fibroblasts ($r = 0.833$) and endothelial cells ($r = 0.822$). This robust performance underscores the capability of DeepDETAILS to handle batch effects and donor variability, ensuring reliable results in real-world scenarios.

A cell-type-specific compendium of regulatory maps in human tissues

Projects such as Roadmap Epigenomics⁴⁶ and ENCODE³ have advanced our understanding of gene regulation across various human tissues and developmental stages. However, their reliance on bulk sequencing results in aggregate signals from heterogeneous cell populations, which can obscure cell-type-specific biology. Rare-cell signals may be masked and opposing regulatory effects across cell types can average out, causing false negatives. To address this limitation, we developed a resource of cell-type-specific regulatory maps across the human body. This resource enhances our ability to understand transcription regulation, annotate noncoding regions of the genome and prioritize noncoding genetic variants associated with diseases. Using DeepDETAILS, we deconvolved bulk ChIP-seq datasets from the ENCODE project, focusing on key histone modifications including H3K4me1, H3K4me3 and H3K27ac in tissues for which snATAC-seq references were available. Additionally, we generated genome-wide maps for transcription initiation at promoters and enhancers across ten human tissues and applied DeepDETAILS to deconvolve their cell-type-specific components (Fig. 5a and Extended Data Fig. 8a). The resulting compendium includes regulatory profiles across 20 organs, 39 tissues and 86 distinct cell types. All datasets are available through our web portal (<https://details.yulab.org>), which supports browsing, visualization and download of deconvolved profiles by assay and biosample category.

To demonstrate the quality of our compendium, we identified regulatory elements exhibiting cell-type-specific activity within individual organs (Methods). These elements were enriched for genetic variants associated with specific biological functions or diseases, aligning with the known roles of the cell types. For instance, in the heart, cardiomyocytes directly influence the electrical properties and ion channel functions that regulate the duration of the interval between the beginning of the Q wave and the end of the T wave on an electrocardiogram (QT interval) during repolarization⁴⁷. Our analysis revealed an enrichment of variants associated with these processes specifically in cardiomyocytes but not in other cardiac cells (Fig. 5b). Cardiac fibroblasts and heart smooth muscle cells have key roles in vascular remodeling, plaque formation and fibrosis, all of which contribute to the development and progression of coronary artery disease^{48,49}. Trends of enrichments were observed specifically in these cell types but not in cardiomyocytes (Fig. 5c). Similar cell-type-specific enrichments were detected for other tissues, including hepatocytes in liver related to complement factor B levels and T cells in lung concerning COVID-19 severity (Extended Data Fig. 8b,c). To further assess the accuracy of the deconvolved tracks, we also validated them against 96 known cell-type-specific markers^{50,51} for 26 cell types. We observed, on average, a 17-fold enrichment of signals in the promoter regions of cell-type-specific markers in the corresponding deconvolved tracks (Fig. 5d). Performance was consistently strong across cell types (Extended Data Fig. 8d), with >97% of marker genes enriched in the expected cell type on average, indicating that DeepDETAILS is broadly robust across the cellular contexts represented in our dataset.

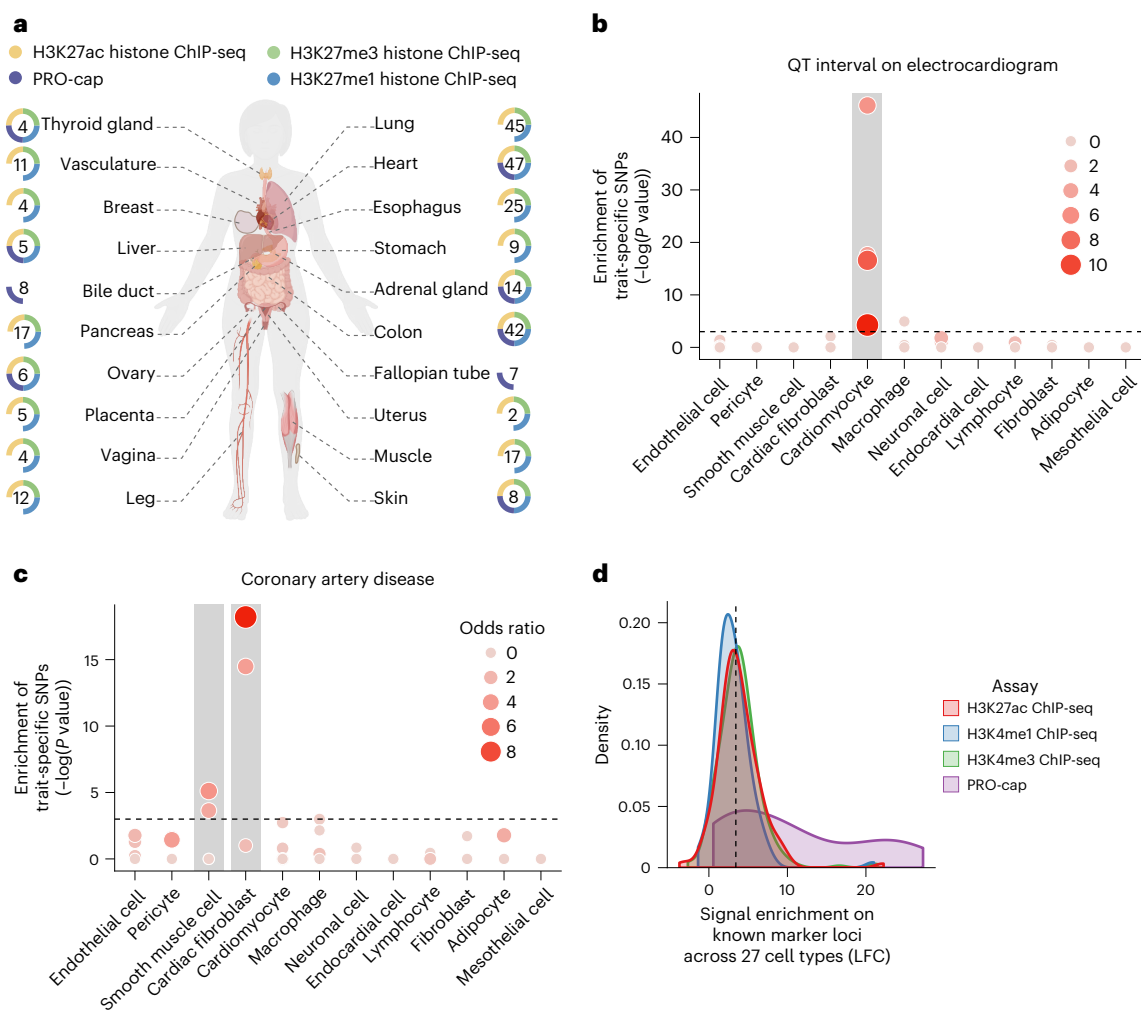


Fig. 5 | Overview of DeepDETAILS compendium. a, Summary of deconvolved datasets in DeepDETAILS compendium. The number in the pies shows the number of cell types deconvolved in the corresponding tissue and the color indicates the availability of the corresponding sequencing modality. **b**, Variants associated with the trait QT interval are enriched in genomic loci that are specifically active in cardiomyocytes. **c**, Variants associated with coronary artery disease are enriched in genomic loci that are specifically active in cardiac

fibroblasts and smooth muscle cells. The horizontal dashed line indicates $-\log(0.05)$. In total, 94 bulk libraries were used for the analysis in **b** and **c**. **d**, Signal enrichment on the promoters of 96 known markers across 26 different cell types (total observations $n = 761$). The dashed line indicates the global median of the log fold change (LFC). Panel **a** created in BioRender; Yu, H. <https://BioRender.com/7jylm07> (2026).

DeepDETAILS compendium facilitates the understanding of disease etiology

High-throughput sequencing has provided unprecedented insights into genetic and molecular mechanisms of disease etiologies⁵². However, bulk libraries capture mixed signals from diverse cell types, which can obscure meaningful patterns and mask cell-type-specific dysregulation. DeepDETAILS compendium addresses these challenges by enabling cell-type-level dissection of regulatory signals in bulk samples, improving our understanding of disease etiology.

Primary sclerosing cholangitis is characterized by chronic inflammation and bile duct scarring (Fig. 6a) but its etiology remains poorly understood. Although genome-wide association studies (GWASs) have identified over 300 progressive cholestatic liver (PSC)-associated variants^{53–55}, pinpointing causal variants remains challenging⁵⁶. To better understand the molecular mechanisms underlying PSC, we used the reconstructed cell-type-specific regulatory maps in DeepDETAILS compendium for transcription initiation, as well as H3K4me1, H3K4me3 and H3K27ac modifications in the human liver's right lobe. We then generated cell-type-specific E–P networks using the activity-by-contact model³⁷ with deconvolved histone modification signals as inputs. The predicted E–P interaction maps were refined by removing enhancer

candidates lacking deconvolved eRNA transcription. Through this approach, we revisited GWAS-identified PSC-associated variants and identified **rs5757584**, located within an enhancer that is specifically active in liver-resident macrophages (Kupffer cells) (Fig. 6b). The inferred E–P interaction map suggested that this enhancer regulates a nearby gene, platelet-derived growth factor subunit B (*PDGFB*). We confirmed that this inference was robust to hyperparameter choice and reference scATAC-seq quality (Extended Data Fig. 9a). To further validate this finding, we compared the transcriptional initiation activity of *PDGFB* across cell types to its expression profile in published scRNA-seq data⁵¹, observing high concordance between our results and the scRNA-seq dataset (Extended Data Fig. 9b). As expected, DeepDETAILS also outperformed statistical deconvolution methods in this cross-modality deconvolution setting (Extended Data Fig. 9c). This E–P interaction was further supported by an independent eQTL study⁵⁸, which demonstrated that the C-to-T mutation negatively impacted the expression of *PDGFB* in lymphocytes (Extended Data Fig. 9d). Previous studies have shown that the secretion of PDGF from macrophages can help suppress the inflammatory response^{59,60}. Therefore, one potential etiology of PSC could be that individuals carrying alternative alleles at **rs5757584** exhibit impaired PDGFB-mediated

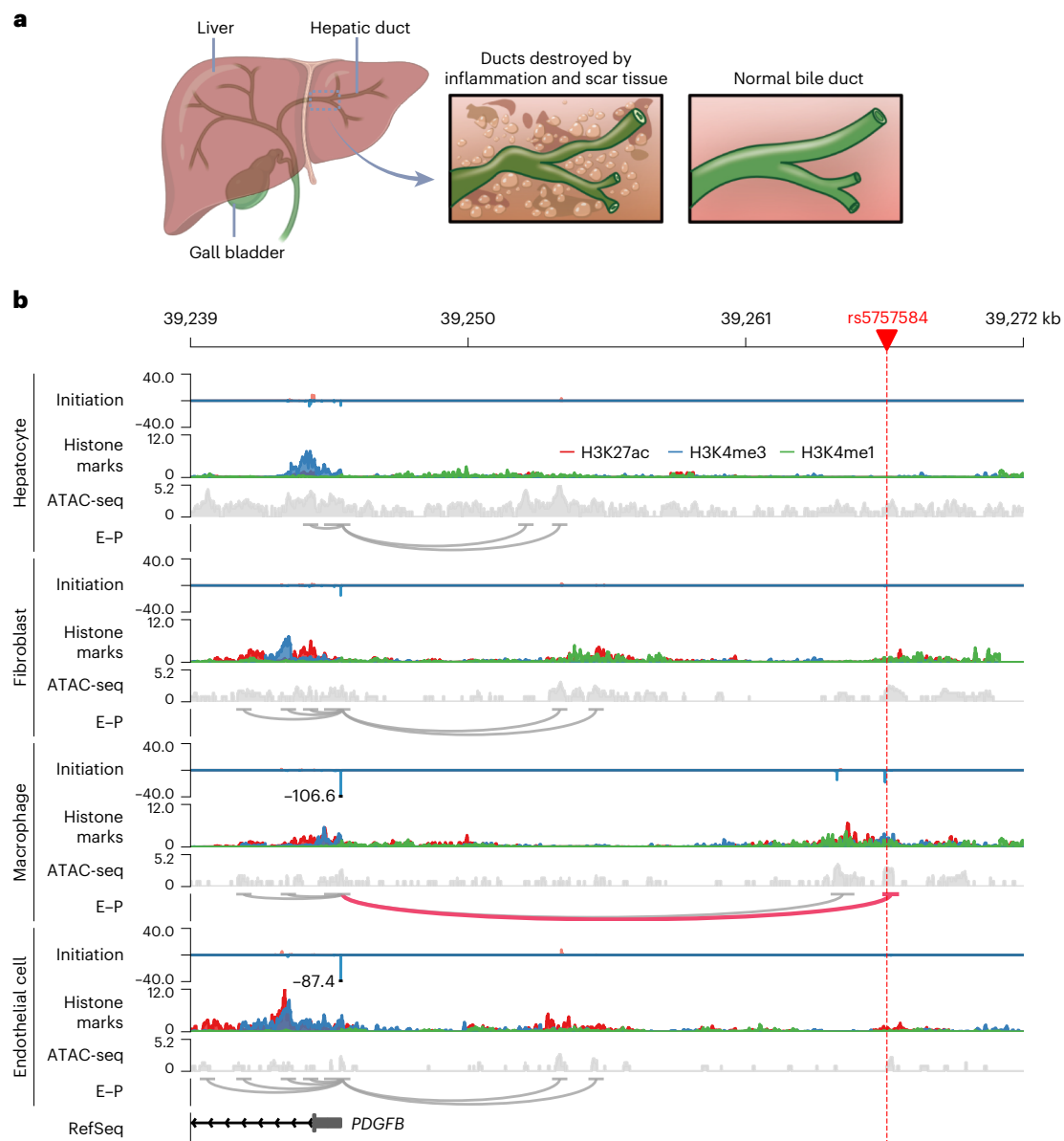


Fig. 6 | DeepDETAILS compendium brings new insight into disease etiology. **a**, Schematic of primary sclerosing cholangitis. **b**, Genome browser view of deconvolved signals at the *PDGFB* locus in the right lobe of human liver. The variant affecting the macrophage-specific enhancer is highlighted at the top,

with the disrupted E-P interaction marked in red. Tracks for transcription initiation and histone modifications are displayed in their original scale. Panel **a** created in BioRender. Yu, H. <https://BioRender.com/92vdfyr> (2026).

immune regulation in Kupffer cells, resulting in a prolonged inflammatory response and increased susceptibility to liver damage. Notably, hepatocytes account for approximately 70% of liver's regulatory signals (Extended Data Fig. 10a), suggesting that cell-type-specific effects in nonhepatocyte populations such as Kupffer cells could be easily overlooked in bulk-level analyses (Extended Data Fig. 10b). Without the cell-type-specific dissection of regulatory processes provided by DeepDETAILS, the functional role of this variant, in terms of both gene regulation and disease etiology, might have remained undetected (Extended Data Fig. 10b).

Discussion

Various high-throughput sequencing assays have advanced our understanding of transcriptional regulation. However, it is difficult to use bulk sequencing assays to identify distinct regulatory patterns in tissue samples composed of diverse cell types. While advances in single-cell sequencing (for example, scRNA-seq and

snATAC-seq) allow high-resolution study of regulatory layers in complex biosamples, numerous other assays are still not applicable at the single-cell level. Current machine learning techniques that use single-cell sequencing data focused on smoothing technical noises in scATAC-seq and snATAC-seq libraries and better representing the clustering structures of the cell population in the sample^{61–65}. When data from another single-cell modality are available, such as scRNA-seq, cross-modality data imputation can be performed to further refine the representations^{66–69}.

DeepDETAILS uses the scATAC-seq and snATAC-seq for a different goal; DeepDETAILS reconstructs regulatory processes in different cell types at base-pair resolution in a bulk sequencing sample when provided with the scATAC-seq or snATAC-seq reference. We showed the broad applicability and robust performance of our method in such a cross-modality deconvolution setting. We further applied DeepDETAILS to deconvolve bulk sequencing libraries across 39 different human tissues and generated cell-type-specific regulatory

maps regarding transcription initiation and histone modifications (H3K4me1, H3K4me3 and H3K27ac). We anticipate the comprehensive and high-resolution cell-type-specific regulatory maps in our portal can help identify molecular mechanisms underlying various pathologies, offering new avenues for therapeutic intervention. For example, the identification of [rs5757584](#) as a variant in a macrophage-specific enhancer affecting *PDGFB* expression highlights how this tool can pinpoint cell-type-specific functional elements that potentially contribute to disease pathogenesis. This finding also underscores the importance of studying regulatory variation at cell-type-specific level and demonstrates the potential of DeepDETAILS to bridge gaps in our understanding of genotype–phenotype relationships.

As a reference-based deconvolution method, the effectiveness of DeepDETAILS depends on the availability, quality and comprehensiveness of the scATAC-seq and snATAC-seq reference datasets. Similar to statistical approaches for intramodality deconvolution, DeepDETAILS cannot resolve signals from cell types absent from the reference. With the advancement and availability of other single-cell modalities, such as single-cell histone modification ChIP-seq, future research could enhance the framework's flexibility to incorporate these as alternative references. Additionally, the current implementation of DeepDETAILS is not optimized for repressive regulatory processes, such as the trimethylation of lysine 27 on histone H3 protein, and modifications on the architecture may be needed. In addition, statistical deconvolution methods usually benefit from a larger number of input bulk samples; thus, it is possible that these tools may perform better when more data are available.

Multimodal foundation models demonstrate potential in generalizing to unobserved cell types and could potentially be repurposed for tasks like deconvolution^{70,71}. However, these models require extensive sequencing data and substantial computational resources; for instance, EpiBERT needed around 740 ATAC-seq datasets. Their effectiveness also depends on the similarity between regulatory motifs in trained and cell types not seen during training. In contrast, the broad applicability of DeepDETAILS across various cell types, regardless of their similarity, enhances its utility in advanced genomic analyses without the need for vast datasets or powerful computing infrastructure.

In conclusion, DeepDETAILS enables cell-type-specific dissection of regulatory landscapes from bulk sequencing data, helping link transcriptional regulation and genetic variation to disease mechanisms.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41587-026-03218-w>.

References

- Soutourina, J. Transcription regulation by the Mediator complex. *Nat. Rev. Mol. Cell Biol.* **19**, 262–274 (2018).
- Malik, S. & Roeder, R. G. Regulation of the RNA polymerase II pre-initiation complex by its associated coactivators. *Nat. Rev. Genet.* **24**, 767–782 (2023).
- ENCODE Project Consortium An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
- Leung, A. K.-Y., Yao, L. & Yu, H. Functional genomic assays to annotate enhancer-promoter interactions genome wide. *Hum. Mol. Genet.* **31**, R97–R104 (2022).
- Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**, 950–953 (2013).
- Tome, J. M., Tippens, N. D. & Lis, J. T. Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. *Nat. Genet.* **50**, 1533–1541 (2018).
- Nojima, T. et al. Mammalian NET-Seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell* **161**, 526–540 (2015).
- Morgan, M. A. J. & Shilatifard, A. Reevaluating the roles of histone-modifying enzymes and their associated chromatin modifications in transcriptional regulation. *Nat. Genet.* **52**, 1271–1281 (2020).
- Core, L. J. et al. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46**, 1311–1320 (2014).
- Kim, T.-K. et al. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).
- Yao, L. et al. A comparison of experimental assays and analytical methods for genome-wide identification of active enhancers. *Nat. Biotechnol.* **40**, 1056–1065 (2022).
- Zhang, J. et al. Comprehensive evaluation of diverse massively parallel reporter assays to functionally characterize human enhancers genome-wide. *Genome Biol.* **26**, 378 (2025).
- Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
- Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
- Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64 (2013).
- Xie, Y. et al. Droplet-based single-cell joint profiling of histone modifications and transcriptomes. *Nat. Struct. Mol. Biol.* **30**, 1428–1433 (2023).
- Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. Methods and applications for single-cell and spatial multi-omics. *Nat. Rev. Genet.* **24**, 494–515 (2023).
- Stuart, T. et al. Nanobody-tethered transposition enables multifactorial chromatin profiling at single-cell resolution. *Nat. Biotechnol.* **41**, 806–812 (2023).
- Mahat, D. B. et al. Single-cell nascent RNA sequencing unveils coordinated global transcription. *Nature* **631**, 216–223 (2024).
- Chen, Y. et al. Deep autoencoder for interpretable tissue-adaptive deconvolution and cell-type-specific gene analysis. *Nat. Commun.* **13**, 6735 (2022).
- Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10**, 380 (2019).
- Jew, B. et al. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nat. Commun.* **11**, 1971 (2020).
- Newman, A. M. et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).
- Chu, T., Wang, Z., Pe'er, D. & Danko, C. G. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat. Cancer* **3**, 505–517 (2022).
- Andrade Barbosa, B. et al. Bayesian log-normal deconvolution for enhanced in silico microdissection of bulk gene expression data. *Nat. Commun.* **12**, 6106 (2021).
- Wang, J., Roeder, K. & Devlin, B. Bayesian estimation of cell type-specific gene expression with prior derived from single-cell data. *Genome Res.* **31**, 1807–1818 (2021).
- Khatry, R., Machart, P. & Bonn, S. DISSECT: deep semi-supervised consistency regularization for accurate cell type fraction and gene expression estimation. *Genome Biol.* **25**, 112 (2024).
- Velmeshev, D. et al. Single-cell genomics identifies cell type-specific molecular changes in autism. *Science* **364**, 685–689 (2019).
- Zheng, S. C., Breeze, C. E., Beck, S. & Teschendorff, A. E. Identification of differentially methylated cell types in epigenome-wide association studies. *Nat. Methods* **15**, 1059–1066 (2018).

30. De Ridder, K., Che, H., Leroy, K. & Thienpont, B. Benchmarking of methods for DNA methylome deconvolution. *Nat. Commun.* **15**, 4134 (2024).
31. Ameen, M. et al. Integrative single-cell analysis of cardiogenesis identifies developmental trajectories and non-coding mutations in congenital heart disease. *Cell* **185**, 4937–4953 (2022).
32. Li, Y. E. et al. A comparative atlas of single-cell chromatin accessibility in the human brain. *Science* **382**, eadf7044 (2023).
33. Zu, S. et al. Single-cell analysis of chromatin accessibility in the adult mouse brain. *Nature* **624**, 378–389 (2023).
34. Novakovsky, G., Dexter, N., Libbrecht, M. W., Wasserman, W. W. & Mostafavi, S. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nat. Rev. Genet.* **24**, 125–137 (2023).
35. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
36. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
37. Dudnyk, K., Cai, D., Shi, C., Xu, J. & Zhou, J. Sequence basis of transcription initiation in the human genome. *Science* **384**, eadj0116 (2024).
38. Avsec, Ž et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
39. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. In *Proc. 34th International Conference on Machine Learning* (eds Precup, D. & Teh, Y. W.) 3145–3153 (PMLR, 2017).
40. Cochran, K. et al. Dissecting the cis-regulatory syntax of transcription initiation with deep learning. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.05.28.596138> (2024).
41. He, A. Y. & Danko, C. G. Dissection of core promoter syntax through single nucleotide resolution modeling of transcription initiation. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.03.13.583868> (2024).
42. Horak, C. E. et al. GATA-1 binding sites mapped in the β -globin locus by using mammalian ChIP-chip analysis. *Proc. Natl Acad. Sci. USA* **99**, 2924–2929 (2002).
43. Wang, L. et al. The multiple roles of interferon regulatory factor family in health and disease. *Signal Transduct. Target. Ther.* **9**, 282 (2024).
44. Zhong, Y., Wan, Y.-W., Pang, K., Chow, L. M. L. & Liu, Z. Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics* **14**, 89 (2013).
45. HuBMAP Consortium The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **574**, 187–192 (2019).
46. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
47. Klabunde, R. E. Cardiac electrophysiology: normal and ischemic ionic currents and the ECG. *Adv. Physiol. Educ.* **41**, 29–37 (2017).
48. Gourdie, R. G., Dimmeler, S. & Kohl, P. Novel therapeutic strategies targeting fibroblasts and fibrosis in heart disease. *Nat. Rev. Drug Discov.* **15**, 620–638 (2016).
49. Wong, D., Turner, A. W. & Miller, C. L. Genetic insights into smooth muscle cell contributions to coronary artery disease. *Arterioscler. Thromb. Vasc. Biol.* **39**, 1006–1017 (2019).
50. Hu, C. et al. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Res.* **51**, D870–D876 (2023).
51. Karlsson, M. et al. A single-cell type transcriptomics map of human tissues. *Sci. Adv.* **7**, eab2169 (2021).
52. Lightbody, G. et al. Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. *Brief. Bioinform.* **20**, 1795–1811 (2019).
53. Ellinghaus, D. et al. Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat. Genet.* **48**, 510–518 (2016).
54. Han, Y. et al. Multitrait genome-wide analyses identify new susceptibility loci and candidate drugs to primary sclerosing cholangitis. *Nat. Commun.* **14**, 1069 (2023).
55. Cerezo, M. et al. The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity. *Nucleic Acids Res.* **53**, D998–D1005 (2025).
56. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
57. Fulco, C. P. et al. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
58. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
59. Spiller, K. L. et al. The role of macrophage phenotype in vascularization of tissue engineering scaffolds. *Biomaterials* **35**, 4477–4488 (2014).
60. Long, H. et al. Macrophages and fibrosis: how resident and infiltrating mononuclear phagocytes account for organ injury, regeneration or atrophy. *Front. Immunol.* **14**, 1194988 (2023).
61. Bravo González-Blas, C. et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* **16**, 397–400 (2019).
62. Xiong, L. et al. Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space. *Nat. Commun.* **13**, 6118 (2022).
63. Yuan, H. & Kelley, D. R. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. *Nat. Methods* **19**, 1088–1096 (2022).
64. Cui, X. et al. Discrete latent embedding of single-cell chromatin accessibility sequencing data for uncovering cell heterogeneity. *Nat. Comput. Sci.* **4**, 346–359 (2024).
65. Hao, M. et al. Large-scale foundation model on single-cell transcriptomics. *Nat. Methods* **21**, 1481–1491 (2024).
66. Wu, K. E., Yost, K. E., Chang, H. Y. & Zou, J. BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. *Proc. Natl Acad. Sci. USA* **118**, e2023070118 (2021).
67. Kalafut, N. C., Huang, X. & Wang, D. Joint variational autoencoders for multimodal imputation and embedding. *Nat. Mach. Intell.* **5**, 631–642 (2023).
68. Tang, X. et al. Explainable multi-task learning for multi-modality biological data analysis. *Nat. Commun.* **14**, 2546 (2023).
69. Cao, Y. et al. scButterfly: a versatile single-cell cross-modality translation method via dual-aligned variational autoencoders. *Nat. Commun.* **15**, 2973 (2024).
70. Gao, Z., Liu, Q., Zeng, W., Jiang, R. & Wong, W. H. EpiGePT: a pretrained transformer-based language model for context-specific human epigenomics. *Genome Biol.* **25**, 310 (2024).
71. Javed, N. et al. A multi-modal transformer for cell type-agnostic regulatory predictions. *Cell Genom.* **5**, 100762 (2025).
72. Moore, J. E. et al. An expanded registry of candidate cis-regulatory elements. *Nature* <https://doi.org/10.1038/s41586-025-09909-9> (2026).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author

self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2026

Methods

Processing sequencing libraries

snATAC-seq libraries for different cell lines and primary cells were processed with 10x Genomics Cell Ranger ATAC 2.1.0 and only fragments from called cells were kept for downstream usage. Run-on sequencing libraries (GRO-cap, PRO-cap, CoPRO and PRO-seq) datasets were processed with pipelines introduced in our previous publication¹¹ and the protocol is available from our GitHub repository. Histone ChIP-seq libraries were processed with the ENCODE ChIP-seq pipeline version 2. All sequencing data and analytical pipelines were managed by BioQueue⁷³.

Simulating bulk tissue datasets for different targeting assays

For run-on libraries (GRO-cap, PRO-cap, CoPRO and PRO-seq) where only the ends of sequencing reads were considered in downstream analysis, we sampled signals from the corresponding bigWig files with a binomial sampler, such that the sampled signal at locus i for cell line k follows $s_{ik} \sim \mathcal{B}(N_{ik}, p_i)$, where N_{ik} is the observed signal value from the original library and p_i is the sampling ratio for cell line k . For histone modification ChIP-seq where the entire reads were considered, we downsampled the aligned reads in the tagAlign files following a uniform distribution $\mathcal{U}(p_i)$. In both cases, the sampled signals for each cell line were merged together and served as the bulk tissue signal tracks. Peaks were called from the simulated bulk libraries using PINTS¹¹, dREG⁷⁴ and MACS2 (ref. 75) for GRO/PRO-cap, 3' of CoPRO/PRO-seq and ChIP-seq, respectively. We surveyed published run-on and histone modification libraries, and used the 25th, 50th and 75th percentiles of their sequencing depth as the criteria of acceptable, good and excellent simulated bulk library (cutoff for each assay in Supplementary Table 3).

Simulating pseudobulk scATAC-seq and snATAC-seq datasets

We used a two-stage hierarchical sampling strategy to simulate scATAC-seq and snATAC-seq libraries for tissues. In the first stage, we uniformly sampled m_k single cells on the basis of their barcodes from the corresponding scATAC-seq and snATAC-seq library for that cell line or primary cell (k in this case); in the second stage, for each sampled cell c_i , we randomly selected n_i reads from the fragments pool of that cell, where $n_i \sim \mathcal{N}(\mu, \sigma)$. We used the (μ, σ) pairs (10,000, 833), (5,000, 333) and (1,000, 167) to simulate excellent, good and acceptable quality, respectively, of scATAC-seq and snATAC-seq libraries, as suggested in a previous study⁷⁶. The final pseudobulk scATAC-seq and snATAC-seq libraries were obtained by pooling all sampled fragments together. Sampling procedures in both stages were performed without replacement.

Architecture of DeepDETAILS

The expanded architectural specifications of DeepDETAILS, including the number of layers, channels, filter sizes and activation functions, are presented in Extended Data Fig. 2. At its core, DeepDETAILS formulates deconvolution for each genomic region as a quasisupervised multibranch prediction problem, in which the observed bulk signal is reconstructed as the sum of latent cell-type-specific signals. Both versions of DeepDETAILS extract sequence embeddings from the 4,096-bp input DNA sequence using dilated convolutional layers (Extended Data Fig. 2c) but differ in sequence encoder depth and complexity; the fused model uses a deeper shared sequence encoder, whereas the seq-only model uses a shallower shared sequence module and relies more heavily on cell-type-specific branches. In the fused version, the sequence embeddings are first adjusted by cell-type-specific gating modules and then combined with local chromatin features extracted from the pseudobulk cell-type-specific scATAC-seq and snATAC-seq profiles for each region and cell type using a shared two-layer GRU. These chromatin features are concatenated with the sequence-derived features and fed into the prediction heads

(Extended Data Fig. 2d,e), enabling predictions that incorporate both sequence and chromatin context.

For each cell type, the resulting features are passed to a pair of heads that predict the signal shape and total count within the center 1,000-bp window; therefore, across all cell types, DeepDETAILS uses K pairs of heads, where K is the number of cell types in the bulk sample (Extended Data Fig. 2d,e), similar to a previous study⁷⁷. Specifically, the shape head produces logits that are normalized across genomic positions by a softmax to represent the positional distribution of signal, whereas the count head predicts the corresponding amplitude after a softplus transformation. The cell-type-specific signal is then obtained as the product of the predicted profile shape and count. For both the fused and seq-only versions, predictions for each cell type are adjusted using a scaling function that zeros out predictions from regions that are inaccessible in specific cell types, as determined from the pseudobulk cell-type-specific tracks of the corresponding scATAC-seq or snATAC-seq library. This accessibility-based scaling function directs each cell type's stack of CNN layers to learn unique and meaningful representations. The scaled predictions from individual cell types are aggregated to form the predicted bulk signal.

Scaling function for cell type k at region r :

$$s(k, r) = \frac{\text{RPM}(k, r)}{\sum_{i=1}^K \text{RPM}(i, r) + \epsilon}$$

where $\text{RPM}(k, r)$ gives the depth-normalized ATAC-seq signals at the center 1-kb region of r in the corresponding cell type k and $\epsilon = 10^{-16}$ to avoid dividing by zero.

During training, DeepDETAILS minimizes the root-mean-squared logarithmic error (r.m.s.l.e.) between the reconstructed bulk signal ($\hat{y}$) and observed bulk signal (y). To prevent the model from collapsing into trivial solutions where multiple branches predict similar outputs, DeepDETAILS implements a redundancy penalty⁷⁸. This penalty discourages excessive correlation between predictions from different regions and improves the overall robustness of the model. Mathematically, the loss function can be expressed as follows:

$$\mathcal{L} = \underbrace{\sqrt{\frac{1}{M} \sum_{i=1}^M \sum_{j=1}^L (\ln(1 + \hat{y}_{i,j}) - \ln(1 + y_{i,j}))^2}}_{\text{r.m.s.l.e.}} + \underbrace{\frac{\alpha}{K} \sum_{k=1}^K \sum_{i=1}^M \sum_{j \neq i}^M c_{k,i,j}^2}_{\text{penalty}}$$

where M is the number of regions in the batch, L is the length of the prediction window, K is the number of cell types, α is a positive constant to balance the r.m.s.l.e. and the penalty and c is the correlation matrix of predictions. Optimization is performed end-to-end by backpropagation from the bulk-level reconstruction loss. Although cell-type-specific target tracks are not observed during training, gradients propagate through the reconstructed bulk signal to each branch-specific shape and count prediction, thereby jointly learning the latent cell-type-specific deconvolution and the bulk reconstruction. In the fused model, gradients further flow through the shared ATAC encoder, the cell-type-specific gating modules and the shared sequence encoder, whereas the accessibility-derived scaling factors are treated as fixed inputs. The default values for M and α are 32 and 1.0, respectively. However, DeepDETAILS generally maintains robust performance across a range of hyperparameter values (Extended Data Fig. 10c,d).

Training and deconvolution are intrinsically coupled in this framework. In each forward pass, the model produced a set of latent cell-type-specific predictions, whose sum defined the reconstructed bulk signal used for optimization. Consequently, the branch-specific outputs after training constitute the final deconvolved cell-type-specific signals, rather than being obtained through a separate post hoc decomposition step. Further information on the design rationale

and mathematical formulation of DeepDETAILS can be found in the Supplementary Notes and Extended Data Fig. 10e,f.

Beyond these shared architectural components, the two DeepDETAILS modes are designed to serve complementary analytical purposes. The fused mode is optimized for quantitative deconvolution performance by integrating both DNA sequence information and cell-type-specific chromatin accessibility profiles, as evidenced in Fig. 3 and Extended Data Fig. 5. This integration enables better performance while maintaining lower computational resource requirements. Accordingly, the fused mode is used as the default for deconvolution because it represents the architecture's highest-performing configuration for cross-deconvolution applications.

By contrast, the seq-only mode is primarily intended for biological interpretation through motif discovery (Fig. 1 and Extended Data Fig. 3). Because it relies exclusively on DNA sequence, it yields cleaner attribution maps when used with 'deepdetails attr', enabling more direct identification of regulatory features underlying deconvolved states. Although this mode is computationally more demanding, it provides important mechanistic insight by revealing candidate TF regulators, thereby facilitating hypothesis generation and guiding downstream experimental validation.

Deconvolution of bulk sequencing libraries using DeepDETAILS

DeepDETAILS requires two inputs: a reference scATAC-seq or snATAC-seq library and a bulk sequencing library for deconvolution. While simulations indicate that DeepDETAILS, trained on one bulk sample, can be applied to other samples comprising the same cell types but differing proportions (Extended Data Fig. 10g,h), we recommend a sample-specific workflow for deconvolving multiple bulk libraries. Specifically, DeepDETAILS should be applied separately to each library to account for potential variations in cell type composition.

Data preprocessing. The deconvolution workflow begins with the 'deepdetails prep-data' command, which requires (1) aligned fragment files from the reference scATAC-seq or snATAC-seq library; (2) cell type annotations for each cellular barcode; and (3) strand-specific signal tracks for the bulk library (for example, bigWig files) and genomic region annotations (for example, peak BED files). This command automatically processes the input data by (1) grouping scATAC-seq fragments by cell type using the provided annotations; (2) generating pseudobulk coverage tracks for each cell type; and (3) storing both the target bulk signal and reference pseudobulk signals in an HDF5 file. To include background regions with G+C content distributions similar to the peak regions, you can sample and add them to the dataset by specifying '--background-sampling-ratio 1.0'. We recommend including these regions when deconvolving histone modification datasets.

Deconvolution procedure. Deconvolution is performed using the 'deepdetails deconv' command with the preprocessed HDF5 file. During training, DeepDETAILS (1) generates cell-type-specific signal predictions; (2) aggregates these predictions into a composite bulk signal; and (3) iteratively optimizes parameters to minimize the discrepancy between the predicted and observed bulk signals. The model checkpoint from the training epoch achieving minimal loss is automatically saved for downstream analysis. For deconvolution tasks, the default mode is fused. Users may optionally use the seq-only mode by specifying the '--seq-only' flag in the deconvolution command. Suggestions regarding mode selection are available in the Supplementary Notes.

Result visualization. For visualization of base-pair-resolution signals, the 'deepdetails build-bw' command exports deconvolved cell-type-specific signals as bigWig files.

Attribution analysis (optional). For sequence attribution analysis (as performed in Fig. 1), the 'deepdetails attr' command can be used. The command takes the compiled dataset and trained model to build attribution tracks in bigWig format.

Comparing DeepDETAILS to supervised learning methods

We trained two sets of Puffin-D³⁷ models and each model was fed by downsampled signals from five sources (cell types) as the authors did in their original manuscript. In one setting, Puffin-D models were trained to predict signals for transcription initiation and pause-release in the other. The combinations of cell lines were identical to what we used in our simulated bulk samples for testing the performance of DeepDETAILS and the formulas are available in Supplementary Table 2. Predictions of Enformer were retrieved from the original publication³⁸ and only predictions for biosamples that were also tested in this study were kept for comparison. Comparisons were evaluated using chromosomal cross-validation.

Benchmark performances on cross-modality deconvolution for statistical methods

Statistical deconvolution methods were used to purify total read counts among regions of interest in 1-kb resolution for each mixed cell type. For CIBERSORTx²³, we applied S-mode batch correction and lowered the filtering threshold for average gene expression to 0, as suggested in the original publication. For BayesPrism²⁴, BLADE²⁵ and DSA⁴⁴, we used the default parameters. For BLADE and DSA, which require specifying signatures for each cell type, we used the signature matrices automatically generated by CIBERSORTx. Invalid predictions (not a number) were replaced by small random values drawn from a distribution that $\ln(X) \sim N(0,1)$ and scaled by 0.1 when calculating correlation coefficients between the predicted values and the ground truth. Additional bulk samples were simulated as statistical methods require multiple bulk samples to run by combining different biological replicates of the corresponding cell lines.

Preflight check

When preparing the data files for running the DeepDETAILS deconvolution workflow, a preflight procedure will be executed by default. This step trims the reference dataset by only keeping cell types consistently present in both the reference and the bulk datasets. We implement this functionality in a three-step manner. First, we build a signature matrix as previously mentioned²³. Second, we use ν -SVR to estimate the fractions of different cell types in the bulk library as suggested by a previous study²³. Third, cell types with very low presence are removed (<3.5%) unless the preflight check is manually bypassed with the option ('--skip-preflight'). The 3.5% cutoff was empirically determined through systematic grid search (Extended Data Fig. 7f), with user-adjustable implementation through the '--preflight-cutoff' parameter. We also noticed that rare cell types present in bulk but absent from reference libraries impose negligible performance impact because of limited signal contribution (Extended Data Fig. 10i,j). While this method may yield lower accuracy in estimating the absolute fraction (median r^2 : 0.567–0.827) compared to intramodality simulations⁷⁹, it remains effective for identifying cell types absent from the bulk sequencing library.

Generation of PRO-cap libraries for human tissues

Human tissue samples were obtained as part of the ENCODE consortium's coordination effort. Solid tissues were pulverized on dry ice using a mortar and pestle and cell lines were cultured in-house according to American Type Culture Collection protocols. A total of 10 million cells were processed per sample. Technical replicates were prepared for tissue samples, whereas cell line experiments were performed using biological replicates. PRO-cap libraries were prepared using a previously described protocol^{5,80,81} with modifications. Briefly, permeabilized

cells, isolated nuclei or pulverized tissues were subjected to nuclear run-on reactions to capture nascent RNA. Total RNA was isolated and split into two rounds of adaptor ligation (with VRA5 + 6N: CCTTGGCAC-CCGAGA AUUCCA rNrNrNrNrNrN, VRA3 + 6N: /5phos/ rNrNrNrNrNrN GATCGTCCGACTGTAGA ACTCTGAAC /3InvdT/) using T4 RNA ligase 1 (New England Biolabs, M0204), followed by reverse transcription using SuperScript III (Invitrogen; 18080044) with custom adaptors. Between adaptor ligation steps, cap-stable selection was performed using CIP (New England Biolabs, M0290) to dephosphorylate uncapped RNAs, thereby preventing ligation to the 5' adaptor and Cap-Clip (CellScript, C-CC15011H) to remove 5' caps from guanylated transcripts, enabling incorporation into the library through 5' adaptor ligation. RNA washes, phenol–chloroform extraction and ethanol precipitation were performed between reactions under RNase-free conditions following manufacturer protocols. PCR amplification was performed using RP1 (AATGATACGGCGACACCGAGATCTACAC GTTCAGAGTTCTACAGTC-CGA) and RPI-n (CAAGCAGAAGACGGCATACGAGAT [index_barcode] GTGACTGGAGTT CCTTGGCACCCGAGAATTCCA) primers, followed by library cleanup and sequencing.

Deconvolution of human tissue samples

For the deconvolution of bulk human tissue samples, we used DeepDETAILS with its default settings. For reference dataset selection, we prioritized snATAC-seq data derived from the same tissue and donor. In cases where such data were unavailable, a composite reference was constructed using all available snATAC-seq datasets for that tissue. The specific reference dataset used for each deconvolution experiment is documented on our web portal. Each deconvolution workflow was performed in duplicate, with average signal values being reported to ensure robustness of the results. For histone ChIP-seq experiments, we merged the deconvolved signals on the two strands and reported binned signals (10 bp) to the portal.

Identification of cell-type-specific regulatory elements and variant enrichment analysis

We identified candidate regulatory elements exhibiting cell-type-specific activity in different tissues or organs using *t*-tests by comparing the depth-normalized signal of a specific assay in a given cell type to that of all other cell types. We controlled for false discovery rate (FDR) using Benjamini–Hochberg's approach and considered regions with an FDR < 0.05 and log fold change > 2 as significant. To assess enrichment of variants, we applied Fisher's exact test for each trait. Within each organ, we identified GWAS variants located in bulk peak regions and categorized them into two groups: those residing in cell-type-specific regions and those in nonspecific regions. We then further classified these variants on the basis of their association with a specific trait and constructed 2 × 2 contingency tables to facilitate odds ratio calculations. Lastly, for each cell type, we calculated the odds ratio of trait-associated variants occurring in cell-type-specific regions compared to nonspecific regions and reported the results. Trait–cell type pairs with adjusted *P* values (Benjamini–Hochberg) smaller than 0.05 were considered significant, indicating a statistically significant enrichment of trait-associated variants in cell-type-specific regions.

Prediction of E–P interaction

An activity-by-contact model was used for predicting E–P interactions⁵⁷. For the cell-type-specific E–P maps, we used a pseudobulk ATAC-seq track for the cells and deconvolved H3K27ac ChIP-seq signal as the measurement of enhancer activity. For the bulk E–P map, we used the entire pseudobulk ATAC-seq and bulk H3K27ac ChIP-seq as the enhancer activity measurement. In both cases, we used Hi-C tracks averaged from multiple human biosamples as the measurement of contact frequency. Predicted interactions were further pruned by removing pairs where the candidate enhancer anchors do not have eRNA transcription (<5 counts per million reads).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All deconvolved tracks are freely available on our portal. Human tissue PRO-cap libraries were deposited to the National Center for Biotechnology Information Gene Expression Omnibus and their accession numbers, together with those of other sequencing datasets retrieved from public databases, are listed in Supplementary Table 1. Cell type annotations for scATAC-seq and snATAC-seq experiments were obtained from the ENCODE portal and the corresponding accession for each deconvolution run is available on our online portal. The data used for the eQTL analysis described in this manuscript were obtained from the GTEx Portal on December 13, 2024. Source data are provided with this paper.

Code availability

The source code of DeepDETAILS is publicly available from GitHub (<https://github.com/haiyuan-yu-lab/DeepDETAILS>). Scripts used to generate the results reported in this study can also be retrieved from GitHub (<https://github.com/haiyuan-yu-lab/DeepDETAILS-analysis>).

References

- Yao, L., Wang, H., Song, Y. & Sui, G. BioQueue: a novel pipeline framework to accelerate bioinformatics analysis. *Bioinformatics* **33**, 3286–3288 (2017).
- Danko, C. G. et al. Identification of active transcriptional regulatory elements from GRO-seq data. *Nat. Methods* **12**, 433–438 (2015).
- Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
- Granja, J. M. et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
- Avsec, Z. et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* **53**, 354–366 (2021).
- Zbontar, J., Jing, L., Misra, I., LeCun, Y. & Deny, S. Barlow twins: self-supervised learning via redundancy reduction. In *Proc. 38th International Conference on Machine Learning* (eds Meila, M. & Zhang, T.) 12310–12320 (PMLR, 2021).
- Maas, Z., Sigauke, R. & Dowell, R. Deconvolution of nascent sequencing data using transcriptional regulatory elements. *Pac. Symp. Biocomput.* **29**, 564–578 (2024).
- Mahat, D. B. et al. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.* **11**, 1455–1476 (2016).
- Judd, J. et al. A rapid, sensitive, scalable method for precision run-on sequencing (PRO-seq). Preprint at *bioRxiv* <https://doi.org/10.1101/2020.05.18.102277> (2020).

Acknowledgements

We thank V. D. Fungal, N. D. Tipples, J. Liang, M. I. Paramo, L. Li, D. Xiong, A. He, Y. Sun and H. Zhu for their helpful discussions. We also thank the ENCODE single-cell working group and the Data Coordination Center, especially A. Wang, B. Doughty, B. Parks, E. Rebboah, E. Kotler, J. M. Schaepe, S. Deshpande, W. Becker, Y. Fu, W. Hou and A. Kundaje, for all their efforts on annotating the cell types. We appreciate the E. Lieberman Aiden lab for sharing the ENCODE Degron lines. This work was supported by grants from the National Institutes of Health (R01AG077899 and R01HG012970 to J.T.L. and H.Y.). This work used Delta at National Center for Supercomputing Applications and Jetstream2 at Indiana University and Cornell University through allocations BIO240120 and

BIO220060 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services and Support program, which is supported by National Science Foundation grants 2138259, 2138286, 2138307, 2137603 and 2138296.

Author contributions

Conceptualization, H.Y. Methodology, L.Y., J.T.L. and H.Y. Software, L.Y. Formal analysis, L.Y. and H.Y. Investigation, L.Y. and A.O. Data curation, L.Y., T.X., Y.Z., M.W., X.P. and A.K.-Y.L. Resources, S.R.S., J.T.L. and H.Y. Writing—original draft, L.Y. Writing—review and editing, A.O., S.R.S., Y.Z., J.Z., J.T.L. and H.Y. Visualization, L.Y., X.P., T.X. and H.Y. Supervision, J.T.L. and H.Y.

Competing interests

The authors declare no competing interests.

Additional information

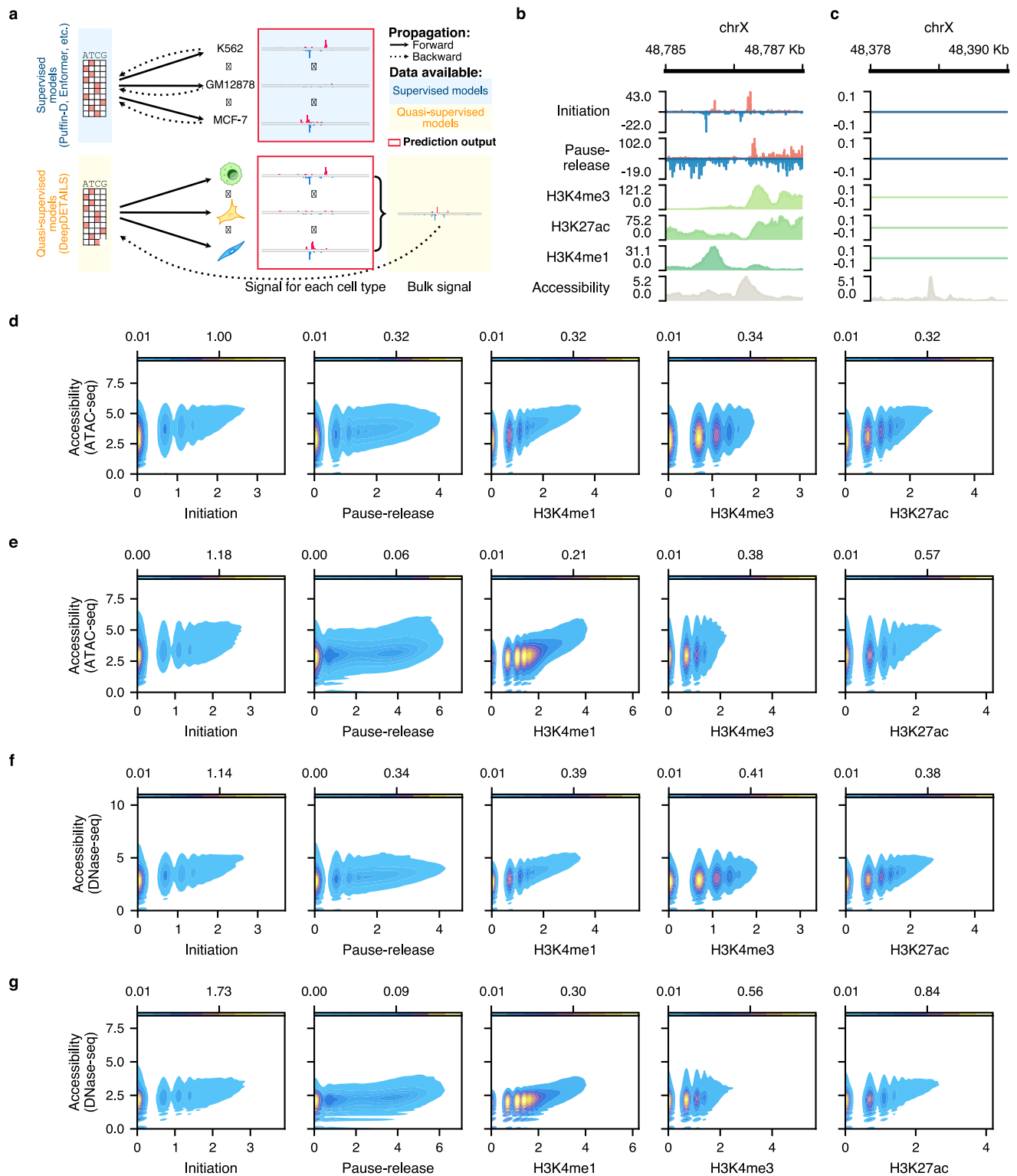
Extended data is available for this paper at <https://doi.org/10.1038/s41587-026-03218-w>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41587-026-03218-w>.

Correspondence and requests for materials should be addressed to Haiyuan Yu.

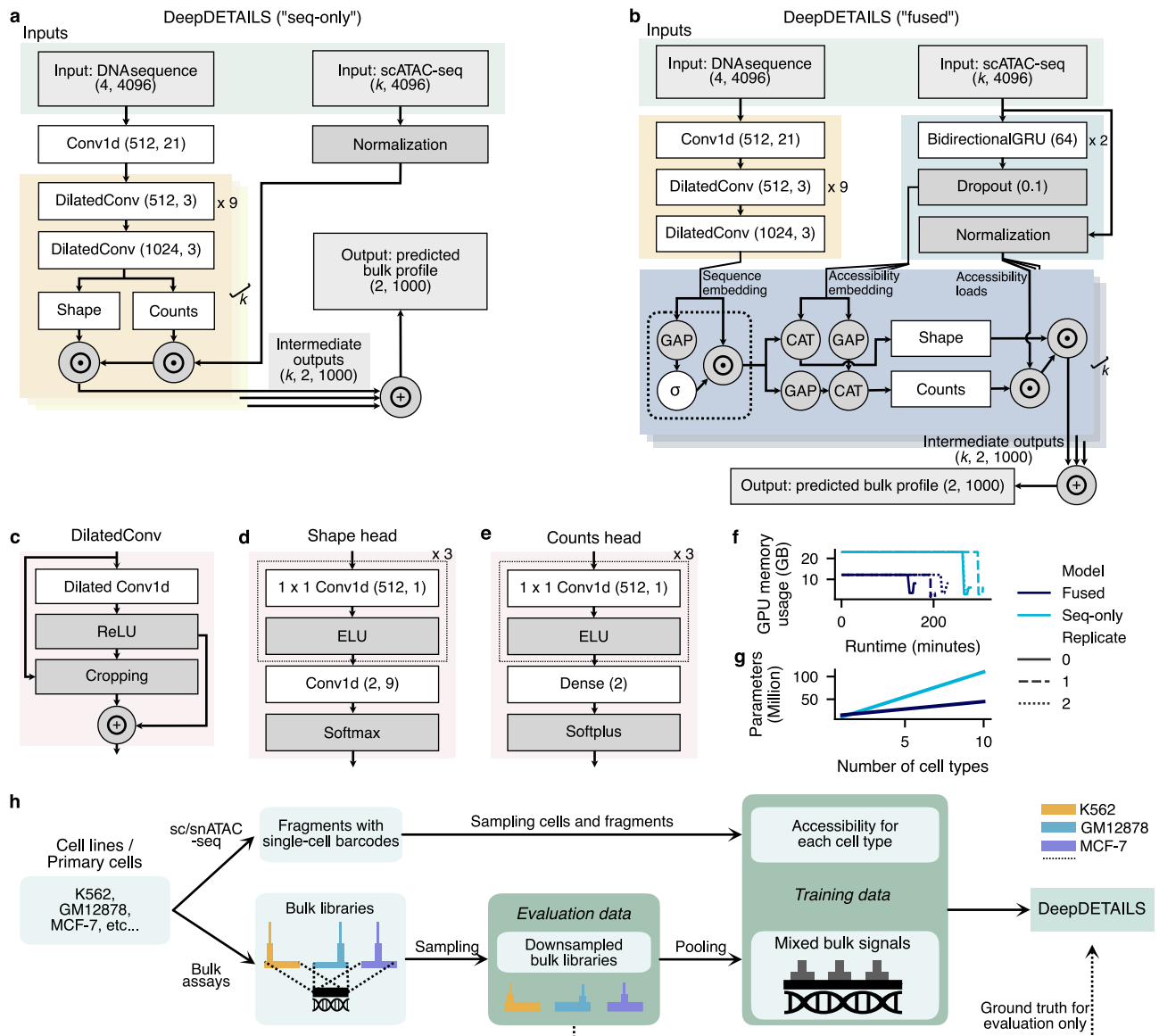
Peer review information *Nature Biotechnology* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.



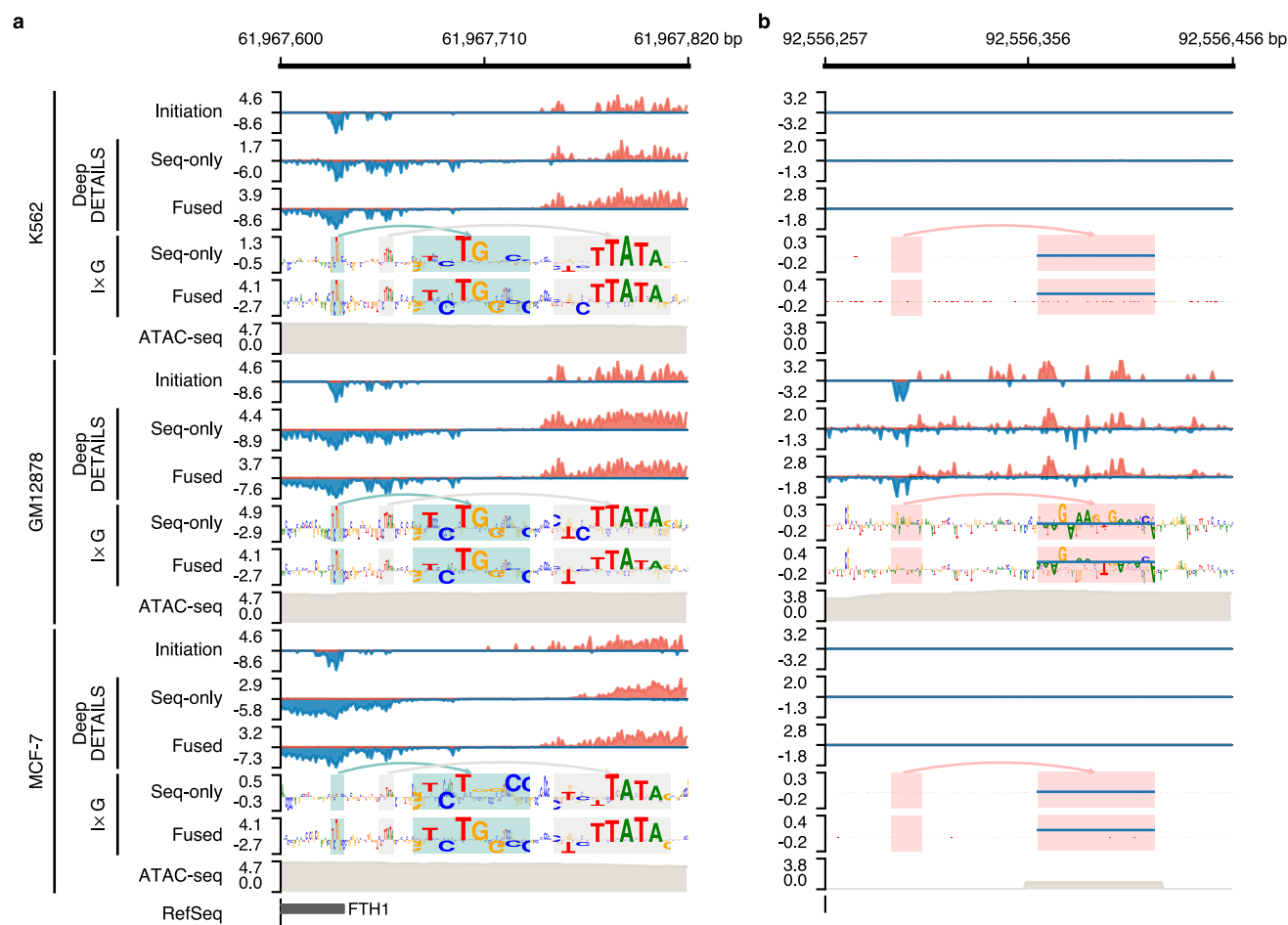
Extended Data Fig. 1 | Accessible chromatin is necessary but not sufficient for many regulatory processes. **a**, Schematic comparison of training strategies used by quasi-supervised methods (for example, DeepDETAILS) and fully supervised models (for example, Puffin-D and Enformer). **b**, Genome browser view of a locus illustrating that accessible chromatin is required for multiple regulatory processes. **c**, Genome browser view of a locus illustrating that

accessible chromatin alone is insufficient to drive regulatory activity. **d–g**, Genome-wide joint distribution analyses comparing chromatin accessibility with other regulatory signals, including transcription initiation, pause release, and histone modifications. Chromatin accessibility was measured by ATAC-seq in **d** and **e**, and by DNase-seq in **f** and **g**. Analyses were performed in K562 cells for **d** and **f**, and in GM12878 cells for **e** and **g**.

**Extended Data Fig. 2 | Expanded characterization of DeepDETAILS.**

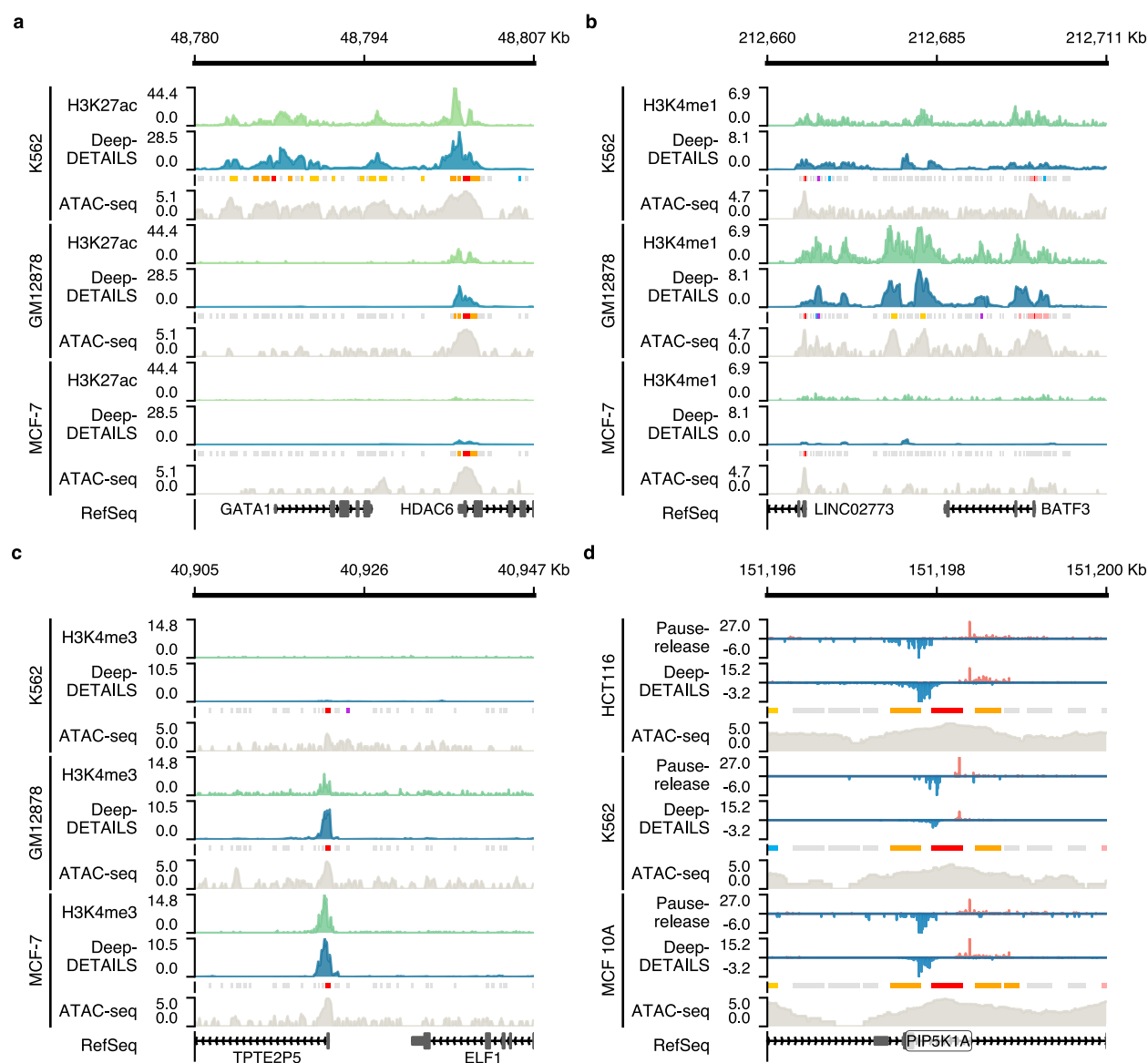
a, Architecture of DeepDETAILS ('seq-only' version). **b**, Architecture of DeepDETAILS ('fused' version). **c–e**, Configuration of dilated convolutional blocks (**c**), heads for predicting the shape of the signal in the 'fused' version (**d**) and total counts in the 'fused' version (**e**). Components with trainable parameters are shown in light colors. Fixed operations are represented in gray. *k*: number of cell types in the bulk sequencing library, GAP: global average pooling, CAT:

concatenation, σ : sigmoid activation, $\odot$: element-wise product, $\oplus$: element-wise summation. **f**, GPU usage and runtime tracking of DeepDETAILS models when applied to the simulated 'tissue' PRO-cap library with three cell types. **g**, The relationship between the number of parameters in DeepDETAILS and the number of cell types in the bulk sample. **h**, Workflow of simulating bulk and reference libraries for evaluating the performance of cross-modality deconvolution.



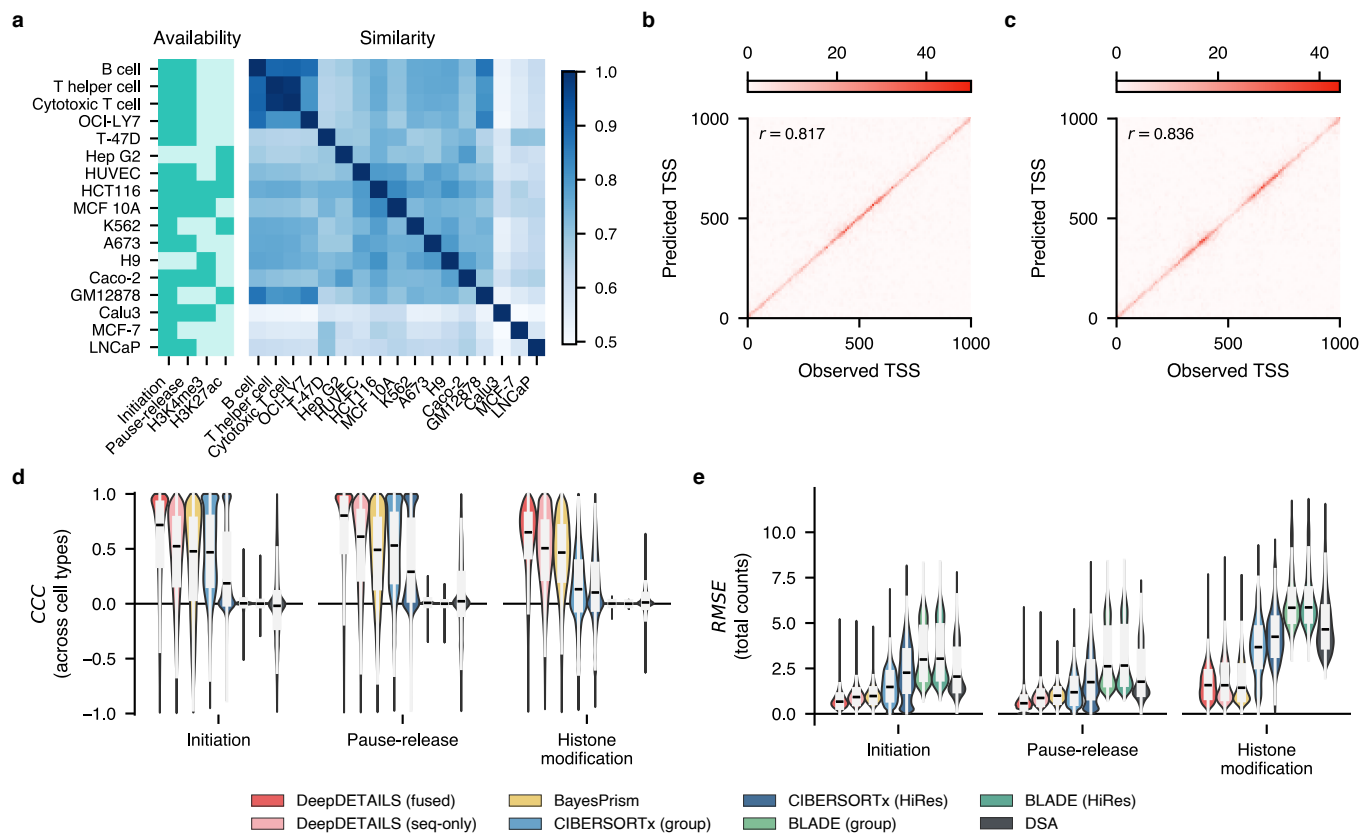
Extended Data Fig. 3 | Expanded sequence attribution analysis. DeepDETAILS deconvolved a simulated bulk PRO-cap library consisting of three cell types: K562, GM12878, and MCF-7. The ground-truth signals were labeled as 'Initiation', while deconvolved signals from the 'seq-only' and 'fused' models were shown separately. Sequence motifs contributing to the prediction in each cell type were shown in the 'I x G' (input x gradient) tracks. In **a**, Initiators and TATA

boxes (consensus sequences BBCABW/WVTGVV and TATAWAW/WTWTATA, respectively) known to be important for transcription initiation in general were highlighted. In **b**, motifs associated with cell-type-specific identity were highlighted (for example, IRF-like motif for GM12878, with the consensus sequence GRAAVWGAAAS). B for C/G/T, R for A/G, V for A/C/G, W for A/T, and S for G/C.



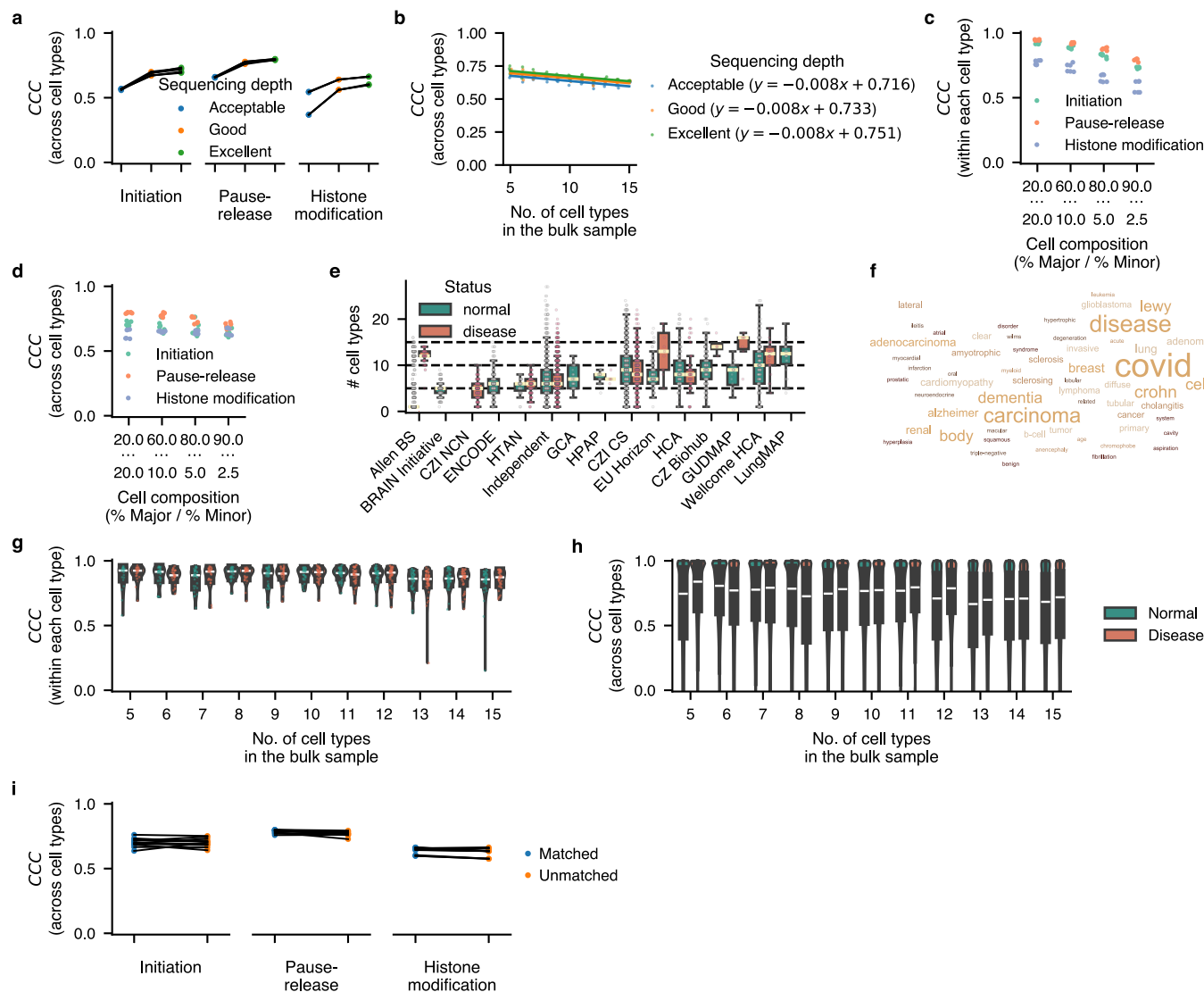
Extended Data Fig. 4 | Expanded demonstration of DeepDETAILS applied across multiple bulk sequencing assays. a–c, Deconvolution of bulk histone modification signals using DeepDETAILS: **a** for H3K27ac, **b** for H3K4me1, and **c** for H3K4me3. **d**, Deconvolution of simulated bulk pause-release profiles

captured by PRO-seq near gene promoters. Candidate regulatory elements such as enhancers (orange boxes) and promoters (red boxes) were annotated as defined in ENCODE cCRE.



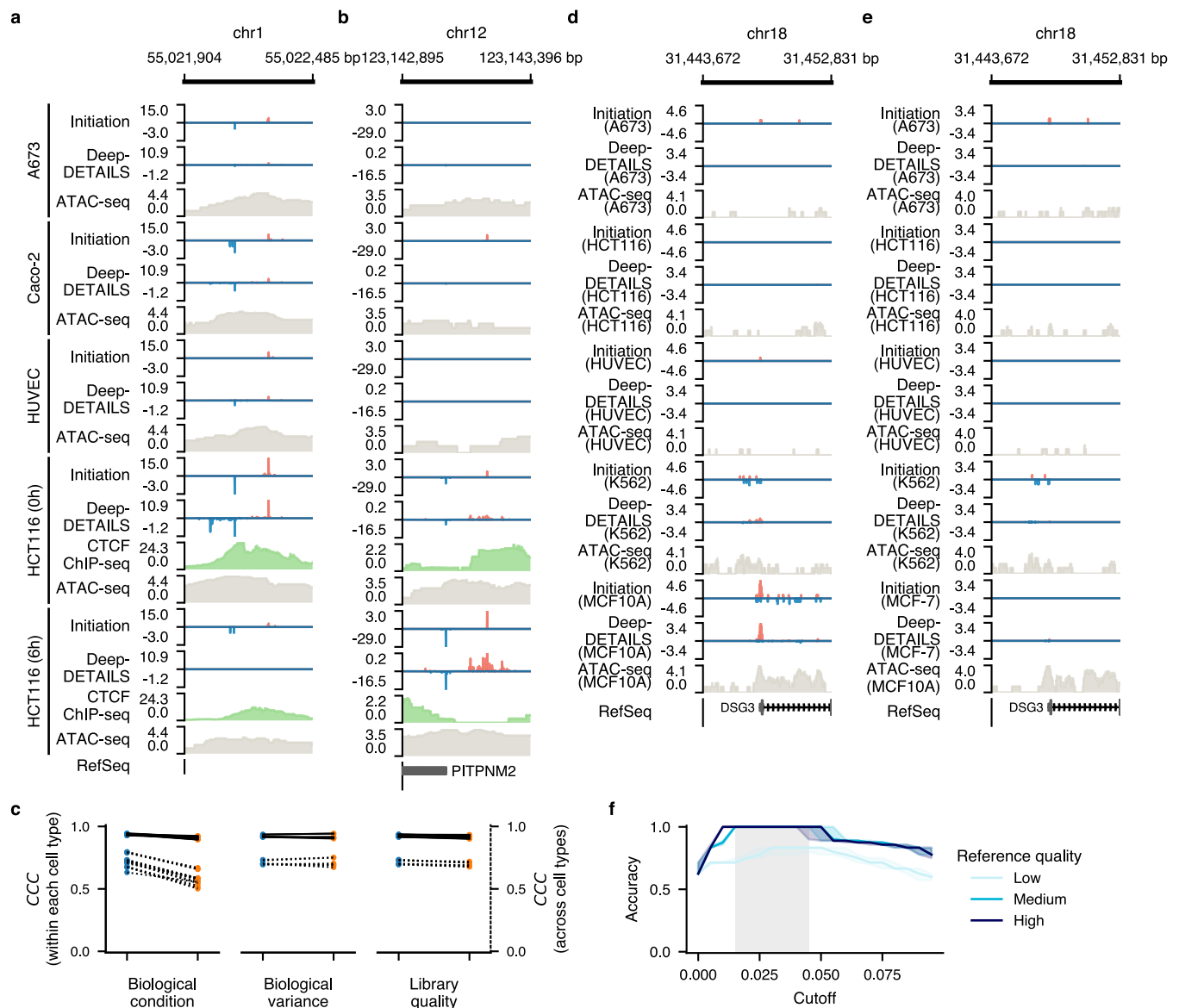
Extended Data Fig. 5 | Extended performance comparison of DeepDETAILS with state-of-the-art methods. **a**, Availability of sequencing libraries for various regulatory steps, with dark color indicating the presence of available libraries (left). Similarity (Pearson's r) in accessible chromatin landscapes across different cell types as measured by snATAC-seq. **b**, Evaluation of consistency (Pearson's r) between Puffin-D reconstructed dominant TSSs and actual dominant TSSs in bulk GRO/PRO-cap libraries. **c**, Evaluation of consistency (Pearson's r) between Puffin-D reconstructed dominant pause sites and actual

dominant pause sites in 3' of bulk CoPRO libraries. For **b** and **c**, we randomly sampled 5000 peak regions for generating the plots. **d** and **e**, Performance comparison between DeepDETAILS and statistical deconvolution tools (1-kb resolution). **d** for across-cell-type evaluation, and **e** for the root-mean-square error evaluation (log-transformed counts). In the box plots the center lines, box limits and whiskers denote the median, upper and lower quartiles and $1.5 \times$ IQR, respectively.



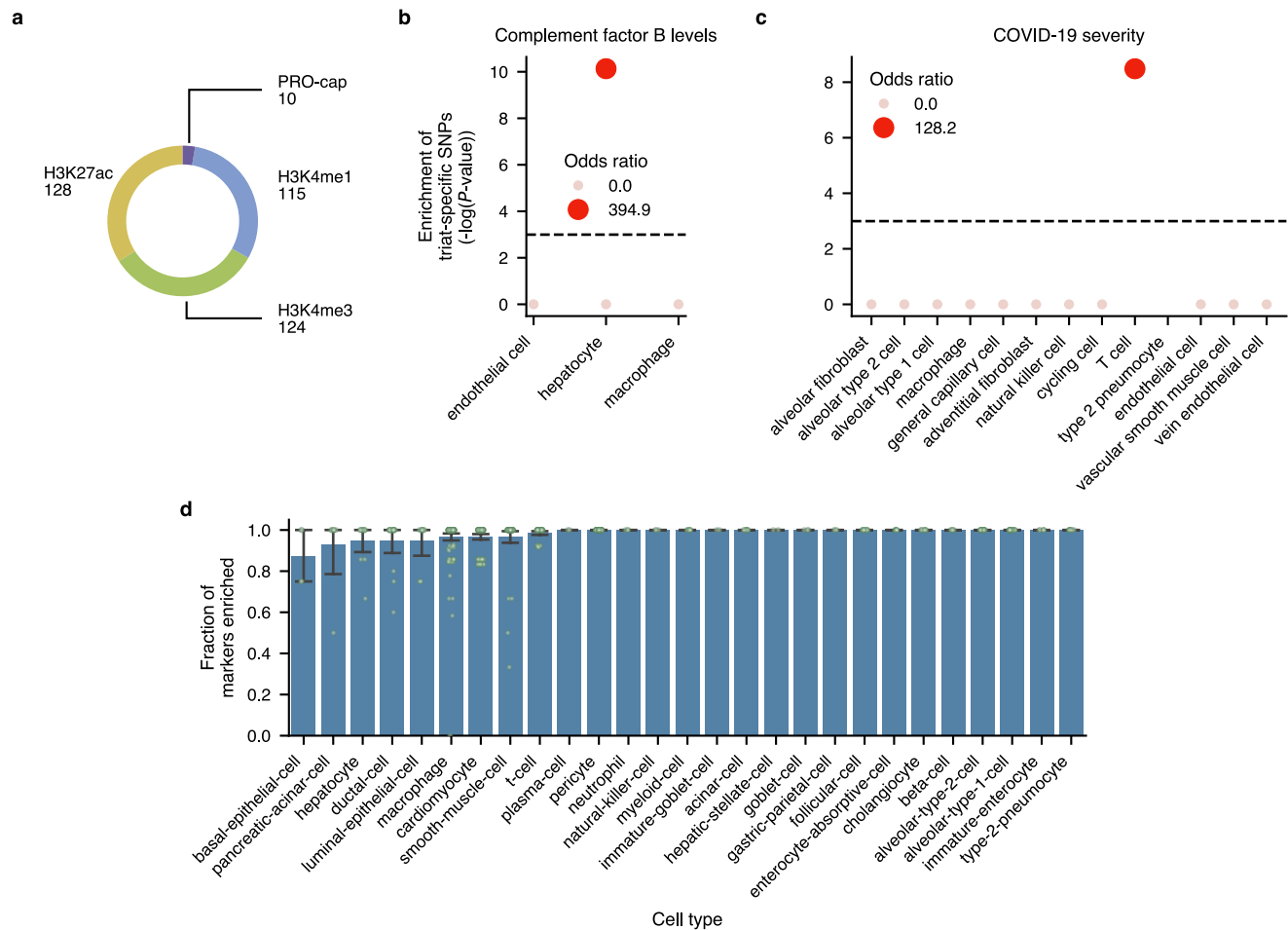
Extended Data Fig. 6 | Extended performance evaluation of DeepDETAILS under challenging cases. **a**, Performance on bulk and reference libraries with varying sequencing depths. For each technique, two bulk-reference pairs were simulated and deconvolved three times ($n = 6$ for each bulk sequencing technique). **b**, Deconvolution of transcription-initiation libraries with increasing numbers of cell types in both bulk and reference libraries. ($n = 3$ for each simulated setting). **c** and **d**, Performance on simulated bulk samples with unbalanced cell-type compositions. X-axis labels indicate the dominant cell-type fraction and the fraction of each of four minor cell types. For each cell composition, six deconvolution runs were conducted per assay. Panel c shows the within-cell-type CCC, and panel d displays the across-cell-type CCC. **e** and

f, Cell-type distributions in real human tissue samples. **e**, Number of annotated cell types per sample ($n = 9790$). **f**, Word cloud of disease terms in the reference, with font size proportional to frequency. **g** and **h**, Performance on simulated bulk samples mimicking real human tissue cell-type distributions, with 5 samples per cell-type-number configuration. Panel g shows the within-cell-type CCC, and panel h displays the across-cell-type CCC. **i**, Performance using unmatched reference libraries, measured by across-cell-type CCC (related to Fig. 4c). For **a**, **b**, **d**, **h**, and **i**, CCC was calculated across cell types as in Fig. 3e. In the box plots (e, g, and h) the center lines, box limits and whiskers denote the median, upper and lower quartiles and $1.5 \times$ IQR, respectively.



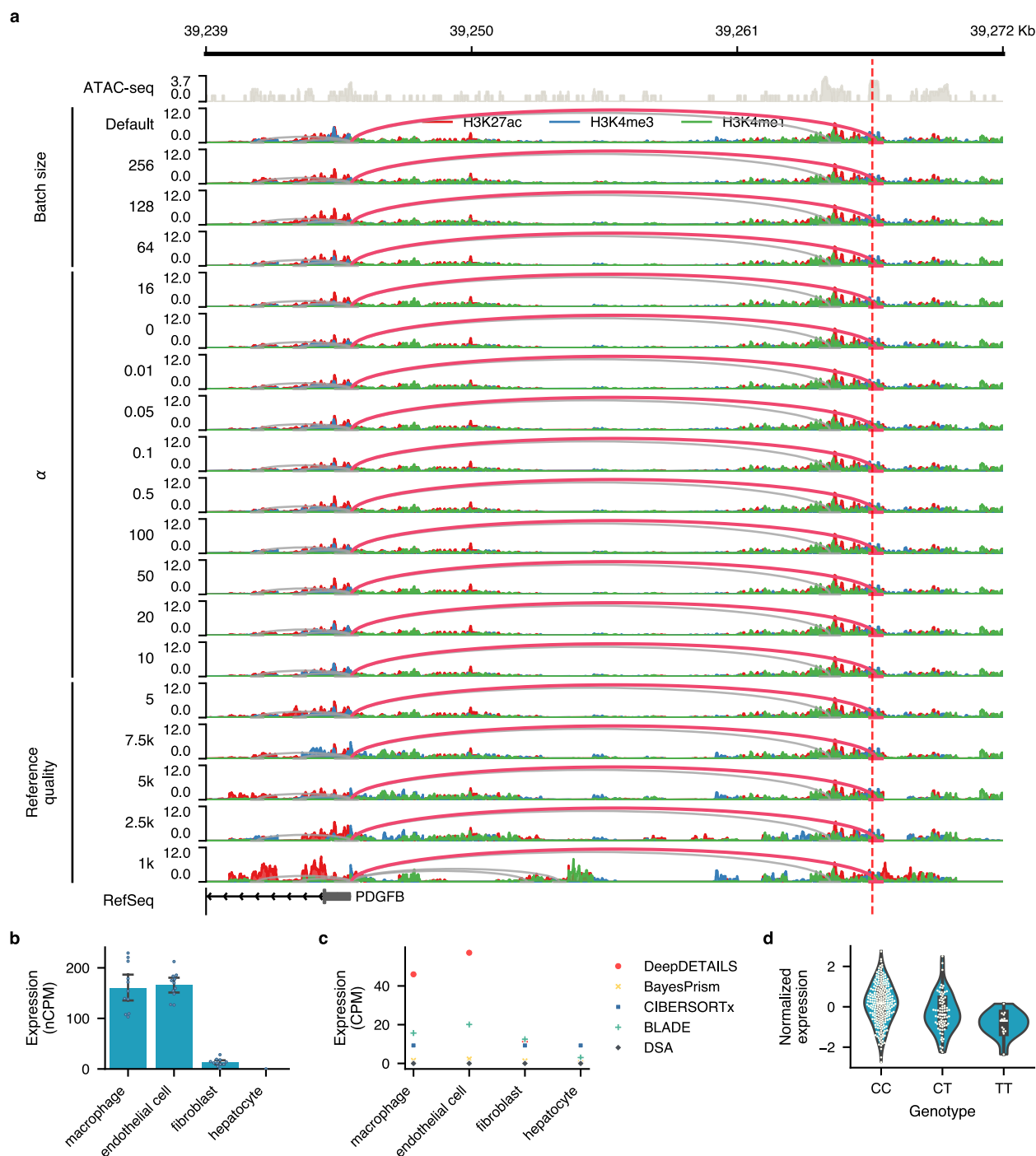
Extended Data Fig. 7 | Extended performance evaluation of DeepDETAILS under out-of-distribution conditions. **a** and **b**, Examples of DeepDETAILS-deconvolved tracks from a simulated bulk sample containing HCT116 cells in different cellular states: control and auxin-induced CTCF degradation. **c**, Evaluation of DeepDETAILS' performances in out-of-distribution scenarios (Supplementary Notes). The left y-axis shows within-cell-type CCC and the right y-axis shows across-cell-type CCC. **d** and **e**, DeepDETAILS

preserves condition-specific information in an out-of-distribution setting. A normal-condition reference (MCF 10A) was used to deconvolve simulated bulk samples containing either normal (MCF 10A, **d**) or cancer (MCF-7, **e**) epithelial cells together with four additional cell types. **f**, Performance of the preflight check step across different cutoffs. Solid lines represent median values, with shaded areas indicating the 95% confidence interval.



Extended Data Fig. 8 | Expanded characterization of DeepDETAILS compendium. **a**, Dataset composition of DeepDETAILS compendium. **b**, Variants associated with the trait complement factor B levels are enriched in genomic loci that are specifically active in liver hepatocytes. **c**, Variants associated with the trait COVID-19 severity are enriched in genomic loci that are specifically active in lung T cells. Two-sided Fisher's exact test was used to derive the *p*-values for

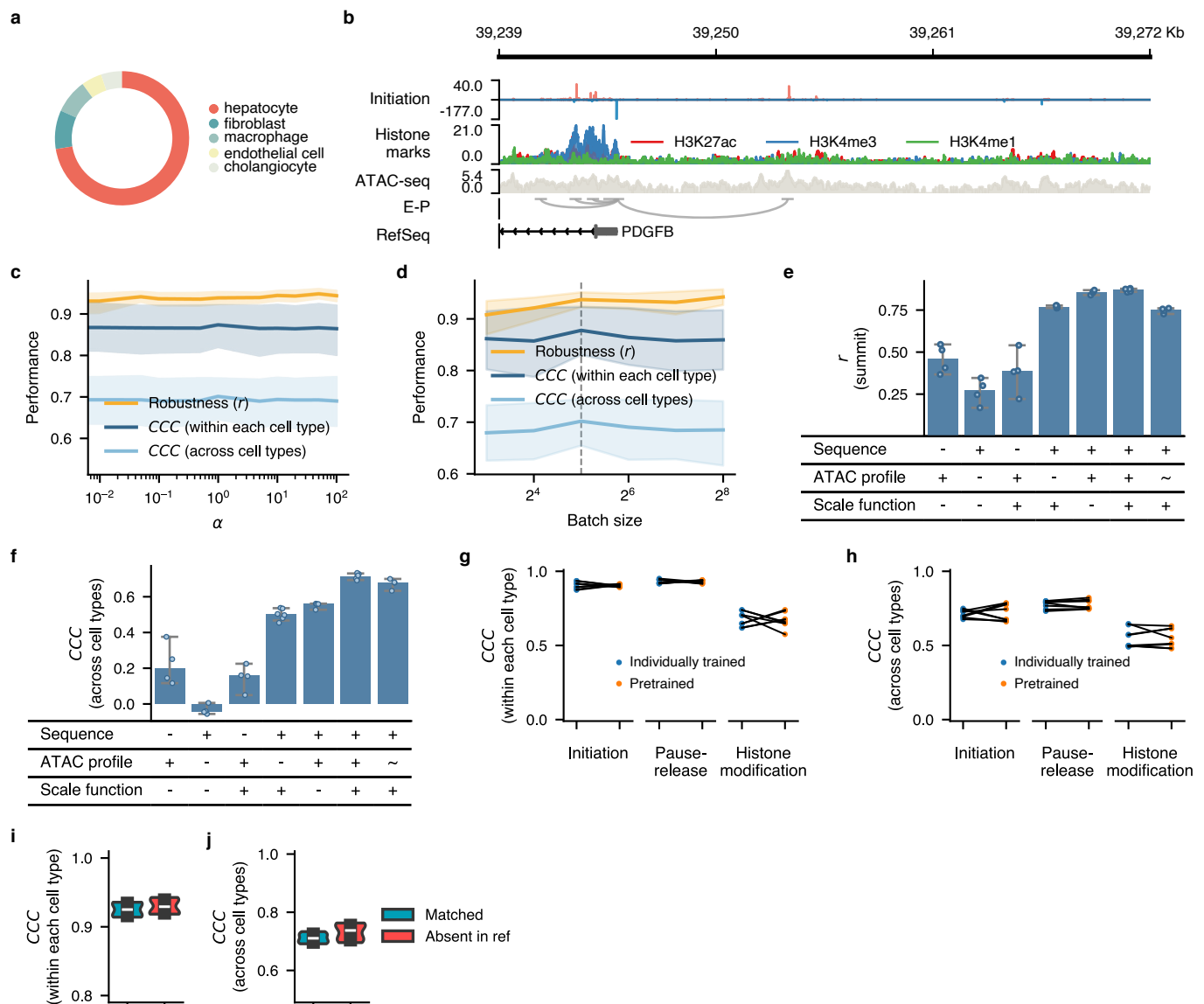
both **b** and **c**. **d**, Enrichment of signals at marker-gene promoters across 26 cell types. Bars denote cell types, and overlaid points show the corresponding raw values from individual deconvolution experiments; the y-axis indicates the fraction of curated marker-gene promoters with enriched signal relative to other deconvolved cell types from the same bulk sample (*n* = 560). Values and error bars represent mean and 95% confidence intervals of mean.



Extended Data Fig. 9 | Extended analysis for DeepDETAILS bringing new insight into disease etiology. a, Evaluation of the robustness of the inferred macrophage-specific enhancer activity to DeepDETAILS hyperparameter settings and the quality of the reference snATAC-seq input, related to Fig. 6b.

b, Gene expression values of *PDGFB* in different cell types measured by scRNA-seq, with sample sizes of $n = 11$ for macrophages, $n = 11$ for endothelial cells, $n = 11$ for fibroblasts, and $n = 1$ for hepatocytes. Values and error bars

represent mean and 95% confidence intervals of mean. **c**, Deconvolved promoter signal for *PDGFB* by different tools. **d**, Impact of the rs5757584 genetic variant on *PDGFB* expression in lymphocytes, stratified by genotype ($n = 237$, 76, and 13 for genotypes CC, CT, and TT, respectively). In the box plot, the center lines, box limits and whiskers denote the median, upper and lower quartiles and $1.5 \times \text{IQR}$, respectively.



Extended Data Fig. 10 | Extended analysis of DeepDETAILS performance and design principles. **a**, Cell composition of a human liver sample from snATAC-seq (5012 annotated cells). Cholangiocyte tracks were excluded from Fig. 6 for consistency because preflight checks indicated their absence in some bulk libraries; tracks from libraries containing cholangiocytes are available on the portal. **b**, Genome browser view of bulk sequencing signals at the *PDGFB* locus in the right lobe of human liver. **c** and **d**, Evaluation of DeepDETAILS' performance using different hyperparameter values: panel c for α and panel d for batch size. Solid lines represent median values, with shaded areas indicating the 95% confidence interval. **e** and **f**, In silico feature-ablation analysis. '+' indicates available features, '-' unavailable features, and '~' shuffled features. $n = 4$ for each condition. Values and error bars represent median and 95% confidence

intervals of median. Panel e for Pearson's r and panel f for across-cell-type CCC. **g** and **h**, Evaluation of DeepDETAILS' performance when a model is pretrained on one sample and applied to other samples with the same set of cell types but in different proportions. For each assay, we generated four pairs of datasets and conducted the deconvolution process twice for each pair. **i** and **j**, Evaluation of DeepDETAILS' performance in scenarios where rare cell types (ranging from 2.5% to 10%) are present in the bulk sample but absent in the reference ($n = 6$ for each condition). Panels g and i show the within-cell-type CCC, and panels h and j display the across-cell-type CCC. In the box plots, the center lines, box limits and whiskers denote the median, upper and lower quartiles and $1.5 \times \text{IQR}$, respectively.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	curl (7.81.0), wget (1.21.2), fasterq-dump (3.0.6)
Data analysis	Data was analyzed using following software tools: BioQueue (2.0.0), fastp (0.20.0), STAR (2.7.1a), samtools (1.11), Python (3.12.8), bedtools (2.31.1), deeptools (3.3.2), numpy (1.26.4), pandas (2.2.3), scipy (1.14.1), matplotlib (3.10.0), seaborn (0.13.2), GenomeViewer (0.1.9), statsmodels (0.14.4), pyfaidx (0.7.0), pybedtools (0.10.0), pyBigWig (0.3.22), biopython (1.84), PINTS (1.1.15), MACS2 (2.2.4), dREG (1.4.0), Django (5.1), bigWigToBedGraph (399), bedGraphToBigWig (399), wiggletools (1.2.10), Puffin-D (34a7278), BayesPrism (2.0), BLADE (56acec2), DSA (1.0), Activity-By-Contact model (c08adc3), R (4.1.2). The source code of DeepDETAILS is available at https://github.com/haiyuan-yu-lab/DeepDETAILS . Analytic pipelines and scripts used to generate results that are reported in this study can be retrieved from https://github.com/haiyuan-yu-lab/DeepDETAILS-analysis .

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

All deconvolved tracks are freely available on our portal. Human tissue PRO-cap libraries have been deposited in NCBI GEO, and their accession numbers, together with those of other sequencing datasets retrieved from public databases (NCBI GEO and the ENCODE Portal), are listed in Supplementary Table 1. Cell type annotations for sc/snATAC-seq experiments were obtained from the ENCODE portal and the corresponding accession for each deconvolution run is available on our online portal. The data used for the eQTL analysis described in this manuscript were obtained from the GTEx Portal on 12/13/2024.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender	In this opportunistic data collection, participants were not selected based on age or sex. Information regarding age and sex is available on the ENCODE portal and our DeepDETAILS portal.
Reporting on race, ethnicity, or other socially relevant groupings	In this opportunistic data collection, participants were not selected based on race, ethnicity, or other socially relevant categories. Information on race, ethnicity, and other socially relevant groupings is available on the ENCODE portal and our DeepDETAILS portal.
Population characteristics	As an opportunistic collection of data, the participants in this study were not selected based on population characteristics.
Recruitment	The ENCODE Data Coordination Center organized the donor recruitment. As this was an opportunistic secondary use study, we did not recruit any participants.
Ethics oversight	Our study utilized data and materials provided by the ENCODE Data Coordination Center. As such, we did not directly recruit or interact with human participants. The ENCODE DCC was responsible for obtaining all necessary ethical approvals and consent from the donors, ensuring compliance with relevant ethical and regulatory standards. Our use of this data adhered to the terms and conditions specified by the ENCODE DCC, and we ensured that all data handling was conducted in line with applicable ethical guidelines.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample sizes calculations were performed -- all data was used where applicable
Data exclusions	No data was excluded from analysis
Replication	All deconvolution-based experiments in the manuscript were performed in duplicate, including all deconvolution experiments displayed on the DeepDETAILS portal. Specifically, for each portal dataset, we ran DeepDETAILS twice, confirmed that both runs converged, and reported the average of the two deconvolved signal tracks. We also assessed the reproducibility of the input sequencing libraries using Pearson's correlation coefficient, with results provided in Supplementary Table 1.
Randomization	Not applicable -- no group-wise experimental testing was performed
Blinding	Not applicable -- no group-wise experimental testing was performed

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.