EAP: a versatile cloud-based platform for comprehensive and interactive analysis of large-scale ChIP/ATAC-seq data sets

Haojie Chen[1,4,#], Tao Huang1[#], Zhijie Guo[1,4,#], Anqin Zheng[1,4,#], Weiran Chen[1], Liangxiao Ma[1], Shiqi Tu[1*], Guangyong Zheng[1,2,*], Yixue Li[1,3,*], Zhen Shao[1,*]

1 Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai, China

2 Shanghai Frontiers Science Center for Chinese Medicine Chemical Biology, Institute of Interdisciplinary Integrative Medicine Research, Shanghai University of Traditional Chinese Medicine, Shanghai, China

3 Guangzhou Laboratory, Guangzhou, China

4 University of Chinese Academy of Sciences, Beijing 100049, China.

* Corresponding authors, Shiqi Tu: tushiqi@picb.ac.cn; Guangyong Zheng: gyzheng@shutcm.edu.cn; Yixue Li: yxli@sibs.ac.cn; Zhen Shao: shaozhen@sinh.ac.cn

# These authors contributed equally to this work

Abstract

Epigenome profiling techniques such as ChIP-seq and ATAC-seq have revolutionized our understanding of gene expression regulation in tissue development and disease progression. However, increasing amount of ChIP/ATAC-seq data poses challenges in computational resources, and the absence of systematic tools for epigenomic analysis underscores the necessity for an efficient analysis platform. To address these issues, we developed the Epigenomic Analysis Platform (EAP, https://www.biosino.org/epigenetics), a scalable cloud-based tool that efficiently analyzes large-scale ChIP/ATAC-seq data sets. EAP employs advanced computational algorithms to derive biologically meaningful insights from heterogeneous datasets and automatically generates publication-ready figures and tabular results, enabling comprehensive epigenomic analysis and data mining in areas like cancer subtyping and therapeutic target discovery.

Introductions

Epigenome profiling plays a crucial role in investigating gene expression regulation in tissue development and disease progression[1,2]. Advanced deep sequencing techniques, including ChIP-seq and ATAC-seq, have enabled us to dissect the epigenetic heterogeneity of developmental cells and disease cohorts by identifying epigenomic alterations and the potential associated transcription factors (TFs)[3,4]. These high-throughput techniques provide valuable insights into the mechanisms that control gene expression and disease progression[5,6].

In recent years, large consortium projects such as ENCODE[7] and TCGA[1] as well as many cancer genomics studies have generated a tremendous amount of ChIP/ATAC-seq data in various normal and cancer cells/tissues, providing opportunities for data mining and a broader understanding of epigenetic regulation of gene expression. However, as the scale of data generation continues to grow, researchers require more computational resources to analyze these large datasets. Additionally, there is a lack of systematic epigenomic analysis tools to explore the huge amount of ChIP/ATAC-seq data deposited in NCBI GEO[8] and CNCB GSA[9]. Despite the availability of a large number of computational tools and methods for ChIP/ATAC-seq data analysis, it remains challenging for experimental biologist to deploy and integrate these tools into workable pipeline, particularly in heterogeneous cohort studies (e.g., those involving patients that may be in different disease states or subtypes) where conventional analysis tools are often inadequate.

To address these gaps, we have developed EAP (Epigenomic Analysis Platform, https://www.biosino.org/epigenetics), a scalable, customizable and interactive ChIP/ATAC-seq data analysis platform based on cloud technology. EAP uses state-of-the-art computational and statistical algorithms to transform ChIP/ATAC-seq data generated from a large panel of samples into biologically meaningful and interpretable results[10-12]. Currently, EAP provides a series of data analytical tools with an interactive web interface to facilitate its use by researchers with limited programming experience, encompassing data preprocessing, supervised differential analysis, differential TF motifs enrichment analysis, differential TF activity analysis, unsupervised hypervariable analysis, clustering analysis, signature genes scoring analysis, etc. All of these analytical tools are specifically developed for modeling and understanding the epigenetic variations among a large number of patient samples or cellular states. The comprehensive epigenomic analysis through EAP can greatly facilitate data mining in different research areas, such as cancer subtyping and therapeutic target discovery.

Results

Overview of EAP architecture and the analysis modules

EAP is primarily tailored for cancer patient cohort studies, necessitating a login for the upload and analysis of sensitive personal information (Supplementary Figure 1a). In order to utilize EAP, users must register an account and request storage space by submitting an application form to the administrator. The input files required for EAP include raw sequencing data in FASTQ format and a metadata file in CSV format, containing details on study design and sample phenotypes. Due to the typically large size of these files, we have developed a dedicated data transfer client tool that supports break-point resumable transfers (Supplementary Figure 1b). Additionally, a md5sum checking procedure has been integrated to ensure the integrity of uploaded files (Supplementary Figure 1c). Currently, EAP offers two analytical modules to transform ChIP/ATAC-seq data from heterogeneous samples into biologically meaningful and interpretable results. EAP leverages private cloud computing technology[13] to establish an analytical framework for processing, interpreting, and visualizing large datasets, ensuring ample computing power for these tasks. Additionally, Docker container technology[14] is utilized to create automated analysis pipelines and tools, ensuring reproducibility of results and streamlining end-to-end bioinformatics analysis for large epigenomic datasets. The cloud computing architecture supports two analytical modules for both basic and advanced epigenomic data analysis in EAP (Figure 1a).

EAP enables efficient and comprehensive large-scale ChIP/ATAC-seq data analysis

The Basic Analysis module is designed and implemented as an automatic pipeline tailored for routine data processing demands, which includes quality control, read alignment, peak calling and read counting. Upon completion of the pipeline, an analysis report (in PDF format), including quality control plots and summary statistic values, will be presented for further study (Figure 1b). NGS data preprocessing is a time-consuming step that demands substantial computational resources. EAP leverages Cloud computing technology to handle large ChIP/ATAC-seq datasets, which can comprise tens or even hundreds of ChIP/ATAC-seq profiles from different samples. Table 1 demonstrates that EAP's basic analysis module can efficiently process and analyze large datasets from various studies.

Table 1. The time cost of data preprocessing using EAP's Basic Analysis module on public data sets with various sample and file sizes

data set	running time	sample size	total file size	source
GBM ATAC-seq	61h45min08s	60	630.6GB	Lu et al.[15]
LUAD H3K27ac ChIP-seq	17h58min36s	42	790.3GB	Yuan et al.[16]
PDAC organoid ATAC-seq	29h19min35s	44	722.6GB	Shi et al.[17]
Thyroid cancer ATAC-seq	120h10min25s	227	3.9TB	Sanghi et al.[18]

GBM: Glioblastoma; LUAD: lung adenocarcinoma; PDAC: Pancreatic ductal adenocarcinoma.

For the Advanced Analysis Module, it encapsulates a comprehensive collection of analytical tools meeting needs of customized ChIP/ATAC-seq data analyses (Figure 1c-d). In practice, users have the flexibility to perform both routine and customized analyses on their data seamlessly, as the input read count table for these advanced analyses is derived from the Basic Analysis module. Furthermore, EAP offers a user-friendly interface that enables users to configure parameters for the advanced analyses, including the number of principal components utilized for sample clustering, the design of the differential analyses, and the specifications for *p*-value/FDR cutoffs, among others (Supplementary Figure 2). Users can choose the specific combination of analytical tools that best suit their research scenarios. For instance, for ChIP/ATAC-seq datasets with clearly defined sample labels, differential analysis can be used to detect the differential signals between samples with different labels, followed by differential TF motif enrichment analysis or differential TF activity analysis to further explore the TFs associated with the differential genome binding or open chromatin sites. This scenario is referred to as supervised analysis (Figure 1c). In contrast, for datasets with no pre-defined sample labels (e.g., those generated from patients with the same disease state/type) or highly sophisticated sample labels (e.g., those covering tens of different cellular states or disease types), users can apply hypervariable analysis to identify hypervariable ChIP/ATAC-seq signals across the samples, which could be then used as features for clustering analysis to dissect the underlying heterogeneity structure among the samples. Then, samples are grouped into different clusters, coupled with signature genes scoring analysis to annotate those clusters based on established gene sets (Figure 1d). Supervised analysis tools could also be applied to the resulting sample clusters to detect binding/open chromatin sites specific to each cluster (Figure 1c-d).

EAP provides interactive data set browser

EAP has been successfully applied to analyze ChIP/ATAC-seq data from many cancer epigenomic studies, including H3K27ac ChIP-seq on LUAD cohort[16], ATAC-seq on NSCLC cohort[19], ATAC-seq on TCGA pan-cancer cohort[1], ATAC-seq on thyroid cancer cohort[18] and ATAC-seq on pancreatic cancer patient derived organoids[17]. These processed data sets are available on the Data Set Browser in EAP, and EAP offers an interactive interface for easy visualization of TF activity scores in each data set. This module allows users to investigate the role of interest of transcriptional regulators in oncogenesis by choosing an appropriate data set (Figure 1e).

Comparison of comprehensiveness of supported ChIP/ATAC-seq features across EAP and other existing analysis platforms/tools

While numerous ChIP/ATAC-seq analysis platforms/tools exist, EAP is designed to provides more flexible and comprehensive down-stream analyses, enabling users to delve deeper into their data and extract biological meaningful insights. Additionally, EAP allows for adding new features in future releases (Table 2). One of the key advantages of EAP is its exceptional ability to handle large-scale ChIP/ATAC-seq datasets ranging from gigabytes to terabytes in size. This capability sets EAP apart from other tools, as it enables users to efficiently analyze and process vast amounts of data.

Table 2. Comparison of comprehensiveness of supported ChIP/ATAC-seq features from currently available analysis platforms/tools

		ChIPseek[20]	PEPATAC[21]	Cistrome[22]	CisGenome[23]	EAP
Data preprocessing	trimming adapter	×	√	√	√	√
	read alignment	×	√	√	√	√
	peaks calling	×	√	√	√	√
	peaks annotation	√	√	×	√	√
	read distribution	×	√	√	√	√
	read counting	×	×	√	×	√
Downstream analysis	motif enrichment analysis	√	√	√	×	√
	differential analysis	×	×	√	×	√
	differential TF motif enrichment	×	×	×	×	√

	analysis					
	hypervariable analysis	×	×	×	×	√
	differential TF activity analysis	×	×	×	×	√
	clustering analysis	×	×	√	×	√
	signature genes score analysis	×	×	×	×	√
	allow for adding new features	×	×	√	×	√
Other	web based analysis tools	√	×	√	×	√
	genome browser visualization	√	√	√	√	×

Case Study 1: ATAC-seq profiles of different tissue types from thyroid cancer patients

To demonstrate the power of EAP with labeled datasets, we collected an ATAC-seq dataset of thyroid cancer[18], comprising 70 primary tumor tissues, 70 patient-match normal tissues and 83 metastatic cancer tissues. We utilized EAP to preprocess the sequencing data and then conducted pairwise differential analysis of different tissue types using count table acquired from the Basic Analysis module. Our analysis revealed that in comparison to the alterations in chromatin accessibility between primary tumor tissues and metastatic cancer tissues, a notably higher number of differentially accessible sites were identified when contrasting primary tumor tissues/metastatic cancer tissues with normal tissues (Figure 2a). We further performed differential TF motif enrichment analysis to detect TF motifs that are differentially enriched in one set of peaks regions relative to another set (e.g., tumor tissue up-regulated peaks vs normal tissue up-regulated peaks). We found that normal thyroid development transcription factors FOXE1 and HHEX were significantly enriched in normal tissue up-regulated sites[24]. However, the potential binding sites for transcription factors of the AP-1 family are more enriched in the up-regulated sites in primary tumors tissue/metastatic cancer tissue, such as FOS, JUN, FOSL2, JDP2, and BATF (Figure 2b)[25].

The AP-1 family is the main target proteins of MAPKs[26], and previous studies have shown that thyroid cancer is driven by MAPK signaling pathway[27]. Therefore, these transcription factors enriched in up-regulated sites in primary tumors tissue/metastatic cancer tissue may be related to the activation of MAPK signaling pathway in thyroid cancer. To verify the activity of MAPK signaling pathway in different tissue types, signature genes activity score analysis tool implemented in EAP was employed to assign the MAPK signaling pathway activity estimates to individual samples. Our results indicated that primary tumor tissues/metastatic cancer tissues have indeed a higher MAPK signaling pathway activity than normal tissues (Figure 2c). Additionally, we performed differential TF activity analysis to identify tissue type specific transcriptional regulators. Consistent with previous findings, TF activity scores of normal thyroid development related transcription factors are higher in normal tissues, while TF activity scores of AP-1 family TFs are increasing in primary tumors tissues/metastatic cancer tissues (Figure 2d-f). It is worth noting that the activity scores of RUNX2 are progressively increasing from patient-match normal tissues to primary tumors tissues, and at last metastatic cancer tissues, which exhibited significantly higher activity (*P*=1.98E-20) (Figure 2g). Enhanced regulation activity of RUNX2 have been reported to be functionally linked to tumor invasion and metastasis in thyroid cancer[28]. Overall, using the combination of EAP differential analysis tools, we found that increasing of AP-1 family TFs regulation ability, up-regulation of MAPK signaling pathway and increased activity of RUNX2 are related to the carcinogenesis and progression of thyroid cancer. These findings demonstrate the functional profiling capacity of EAP in capturing epigenetic alterations and the associated transcription factors in labeled data set.

Case Study 2: ATAC-seq profiles of patient-derived pancreatic cancer organoids

However, in most cancer studies, due to the heterogeneity among patients and the complexity or lack of clinical pathological information, it is difficult to use existing information as sample labels for down-stream analysis. Consequently, EAP provide an unsupervised analysis strategy for analyzing label-free dataset. We applied EAP to reanalyze an ATAC-seq dataset containing 44 patient-derived pancreatic cancer organoids (Figure 3a)[17]. Four different pancreatic ductal adenocarcinoma (PDAC) subgroups were identified based on previously defined signature gene sets[29,30], including classical progenitor like subgroups (C1), basal-like subgroups (C2), classical immunogenic like subgroups (C3) and aberrantly differentiated endocrine and exocrine (ADEX) like subgroups (C4 and C5) (Figure 3b-c). According to basal subtype signature gene scores, we divided the four subgroups into basal-like and non-basal samples, basal-like PDAC showed significantly worse prognosis than those in non-basal patients (*P*=0.045), which is in line with previous studies (Figure 3d)[5,31]. To further elucidate differences in transcriptional regulators between different subgroups, we identified subgroup-specific transcription factors using the differential TF activity analysis tool. GATA6, HNF1A, and HNF1B, which are critical for normal pancreatic development and differentiation, were enriched in classical progenitor like subgroups[32-34]. In the classical immunogenic-like subgroup, HNF4A and HNF4G were identified, and these transcription factors were reported not only to be associated with the classical pancreatic cancer subtype, but also to play an important role in the invasion of pancreatic cancer[35]. RUNX1, RUNX2 and RUNX3 were significantly enriched in the basal-like subgroup, and the high expression of these transcription factors have been reported to promote pancreatic cancer cell migration, invasion and drug resistance, which can be used as a potential target for pancreatic cancer treatment[36,37]. Transcription factors related to pancreatic endocrine or neuroendocrine, such as NEUROD2 and ASCL1[38], were significantly enriched in the ADEX like subgroup. In addition, Rbpj1 is critical in the differentiation of pancreatic acinar cells and it was also enriched in ADEX like subgroup(Figure 3e)[39]. These results indicate that both endocrine and exocrine transcription factors are highly activated in this subgroup, rather than one of the endocrine or exocrine-related transcription factors, which is common in normal pancreas development. Overall, these findings exemplified the utility of EAP in revealing both cancer epigenetic subtypes and subtype specific transcriptional regulators in a label-free data set.

Discussions

In summary, EAP presents a revolutionary data analysis platform to support large-size epigenomic dataset analyses based on the cloud combining technology. We demonstrated the practicability and high-efficiency of EAP in dissecting cancer epigenetic heterogeneity, which can be used not only for mining key transcriptional regulators of different cancer pathological subtypes, but also for cancer epigenetic subtyping, and further revealing subtype associated transcription factors.

Furthermore, EAP offers interactive data analysis, allowing end users to iteratively modify analytical parameters to enhance the identification of biologically significant results. This feature is particularly crucial for cancer epigenomic data analysis, where a standardized analytical pipeline may not be suitable or preferred.

EAP has been successfully applied to analyze ChIP/ATAC-seq data from many cancer epigenomic studies, which comprise tens or even hundreds of ChIP/ATAC-seq profiles from different individuals. The processed data sets are available on the Data Set Browser in EAP, and EAP offers an interactive interface for easy visualization of TF activity scores in each data set. It is worth noting that as we applied this module on datasets from multiple cancer types, RUNX family TFs exhibited higher regulation activity in both progressive subtypes and metastatic tissues (Figure 3f and Supplementary figure 3), which indicate RUNX family TFs might as candidate multi-cancer progression related regulators and may be a potential target for cancer therapy.

We anticipate that EAP can be used for more large-scale cancer epigenomic data analyses, leading to more interesting and meaningful biological or clinical discoveries, which broaden our understanding of cancer carcinogenesis and progression related epigenetic machinery.

Methods

Implementation

The EAP web server runs on a private cloud system equipped with 5,654 GB of memory, 824 CPUs, and 10 TB of storage space. Docker technology was utilized to create images containing necessary software and packages for ChIP/ATAC-seq data analysis, such as *FastQC* (v0.11.9) for read quality control, *Bowtie* (v1.2.0) for read alignment, and *MACS* (v1.4.2) for peak calling. Analysis workflows, including data preprocessing pipelines and downstream analysis tools, were defined in YAML files specifying configuration parameters and commands to be executed. Subsequently, graphical user interfaces (GUIs) for the data preprocessing pipeline and downstream analysis tools were deployed on the web server. The user-friendly GUI was designed to allow non-technical users to customize analysis parameters (e.g., adjusted *p*-value cutoff, number of clusters, variables of interest for differential analysis) and conduct various ChIP/ATAC-seq data analyses.

EAP implements login authentication as a crucial measure for data security and storage. Users are required to register an account (https://www.biosino.org/epigenetics/#/user/register) and request storage space by submitting an application form to the administrator (application form was available at https://github.com/haojiechen94/EAP/blob/main/doc/File_1_Storage_space_application_form.xlsx). Once users have an account and allocated storage space, they can upload raw sequencing data to their designated storage area and utilize the analysis tools integrated within EAP. Instructions and help menus are readily accessible on the Home page to guide users through the platform (help manus was also available at https://github.com/haojiechen94/EAP/blob/main/doc/Help%20document.pdf). Additionally, we also provide video tutorials that visually demonstrate the processes and make it easier for users to learn and apply the procedures effectively (Supplementary movie 1, Supplementary movie 2 and Supplementary movie 3, https://github.com/haojiechen94/EAP/tree/main/video_tutorials). Given the typically large size of uploaded files, a data transfer client was developed to support break-point resumable transfers. Users are advised to utilize this client for uploading raw sequencing data to their allocated storage space within EAP.

Basic Analysis module: a standardized ChIP/ATAC-seq data preprocessing pipeline

Input:

The input for the data analysis process includes raw sequencing data in FASTQ format and metadata containing descriptions of the raw sequencing data and phenotypic information for each sample. A metadata template can be accessed on the Home page of the EAP platform (metadata template was also available at https://github.com/haojiechen94/EAP/blob/main/doc/File_2_metadata.csv).

Data Processing:

Firstly, *FastQC* (v0.11.9) is utilized for quality control of read bases and *Trim-galore* (v0.6.7) is employed for trimming adapters and removing low-quality bases from reads. ChIP/ATAC-seq reads are aligned to the user-specified reference genome, which can be downloaded from the UCSC genome browser (https://hgdownload.soe.ucsc.edu/downloads.html), using *Bowtie* (v1.2.0). Then, PCR duplicates are removed based on genomic position and remaining reads are used for peak calling with *MACS* (v1.4.2). Finally, a count table is generated using *MAnorm2-utils* (v1.0.0).

Output:

The analysis generates three count tables where genomic regions are represented in rows and samples in columns.

- One count table (NA_profile_bins.xls) is prepared for differential analysis.
- Two count tables (proximal_peak_regions_2000bp.txt and distal_peak_regions_2000bp.txt) are prepared for hypervariable analysis.
- A summary report is produced, encompassing quality control assessments for bases, reads mapping, and peaks calling, as well as genomic annotations of peaks and motif enrichment within peaks.

Advanced Analysis module: a comprehensive collection of ChIP/ATAC-seq data analysis tools

Differential Analysis

Input:

The input for the differential analysis module includes count tables (NA_profile_bins.xls) generated from the output of the Data Preprocessing module and metadata.

Data Processing:

Raw count tables are normalized using *MAnorm2* to account for differences in library size and correct MA trended bias. Then differential analysis is conducted based on the user-specified variable of interest.

Output:

The output of the analysis includes:

- A table presenting the results of the differential analysis.
- MVC plot illustrating the global mean-variance trend.
- MA plot highlighting significantly differentially enriched or accessible peaks.

Differential TF Motif Enrichment Analysis

Input:

The input for the differential TF motif enrichment analysis module consists of the result table from the previous step of the differential analysis.

Data Processing:

A user-selected adjusted *p*-value cutoff is applied to define significantly differentially enriched or accessible peaks (DEPs or DAPs). Then *motifscanR* tool is utilized to scan for transcription factor (TF) motif occurrences in these DEPs/DAPs. Finally, Fisher's exact test is employed to determine if a TF motif is significantly enriched in one set of peak regions relative to another set.

Output:

The output of the analysis includes:

- A table presenting the results of the differential motif enrichment analysis.
- Enriched/depleted *p*-values indicating the extent to which a TF motif is over/under-represented in a peak set relative to another peak set.
- A volcano plot highlighting significantly differentially enriched TF motifs.

Hypervariable Analysis

Input:

The input for the hypervariable analysis module includes count tables (proximal_peak_regions_2000bp.txt and distal_peak_regions_2000bp.txt) obtained from the output of the Basic Analysis module and metadata.

Data Processing:

MA normalization is performed in pseudo-reference mode. *HyperChIP* is applied to evaluate the significance of signal variability. Peak regions with an adjusted *p*-value below 0.001 (default) are identified as hypervariable regions. Then [rincipal component analysis (PCA) is conducted based on signals within these variable regions and samples are visualized in two-dimensional PCA space or *t*-SNE space.

Output:

The output of the hypervariable analysis tool includes:

- Hypervariable analysis results exported in both TXT and RData formats.

- MVC plots illustrating the global mean-variance trend in proximal and distal regions, with significant hypervariable regions marked in red.
- Scatter plots displaying the PCA results or t-SNE dimension reduction results based on signals in hypervariable regions.

Clustering Analysis

Input:

The input for the clustering analysis includes hypervariable analysis results in RData format and metadata.

Data Processing:

Principal component analysis (PCA) is conducted on signals within hypervariable regions and the top-ranked principal components (PCs) are utilized to calculate the Euclidean distance between each pair of samples. Then hierarchical clustering is applied to group samples into clusters based on the similarity of their signal patterns in hypervariable regions (HVRs).

Output:

The output of the clustering analysis tool includes:

- A new metadata containing the clustering results, which can be utilized for further analyses.
- A cluster heatmap displaying the hierarchical clustering result, providing a visual representation of sample similarities and clustering patterns.

Signature genes score analysis

Input:

Hypervariable analysis result of proximal regions in RData format, genes to proximal regions links and a list of genes of interest in GMT format.

Data processing:

Given a gene set of interest, it is usually more desirable to summarized the expression level of that gene set using a single integrated score. This tool standardizes the ChIP/ATAC-seq signals in the proximal regions within a given dataset by *z*-score transformation. Then summarizes resulting scores of those proximal regions linked to the genes of interest, minus the mean of z-scores of all proximal regions as negative control.

Output:

The output of the Signature genes score analysis tool includes:

- A boxplot shows the distribution of signature scores among different clusters. This analysis provides visualization to improve interpretation of the clustering results. For example, users can annotate the cluster based on these signature genes scores.

Differential TF Activity Analysis

Input:

Hypervariable analysis results in RData format and metadata.

Data processing:

Peak regions with adjusted *p*-value below 0.001 (default) were defined as hypervariable regions. *motifscanR* was then utilized to conduct motif scanning on these genomic regions, aggregating TF motif-associated signals in each sample to generate a score representing TF regulatory activity. Following this, a *t*-test was employed to identify TFs associated with the user-specified variable of interest.

Output:

The outputs in this analysis including a table of TF activities in each sample, a table of *t*-statistic of the association test and plots for dimension reduction visualization of samples and the activities of user specified TFs.

Evaluation

We showcase application of EAP with two distinct analysis scenarios: (i) In data set with clearly defined labels. In this analysis scenarios, we reanalyzed an ATAC-seq data set of thyroid cancer, which included 70 primary tumor tissues, 70 patient-match normal tissues and 83 metastatic cancer tissues. Using EAP, we investigated chromatin accessibility alterations during cancer progression and identified potential associated transcriptional regulators. (ii) In data sets without pre-defined labels. For this case study, we obtained ATAC-seq data on 44 patient-derived pancreatic cancer organoids. First, we perform sample clustering based on chromatin accessibility heterogeneity and define subtypes using established signature gene sets. Subsequently, we identified key regulators specific to each subtype.

Data availability

This study made use of multiple publicly available datasets. For Case Study 1, the thyroid cancer ATAC-seq was obtained from GEO, under accession number GSE162515. For Case Study 2, the PDAC organoid ATAC-seq data was obtained from GSA, under accession number HRA002013. The LUAD H3K27ac ChIP-seq data was downloaded from EGA, under accession number EGAD00001007066. The GBM ATAC-seq data was obtained from GEO, under accession number GSE163853. The NSCLC ATAC-seq data was obtained from West China Hospital (https://pms.cd120.com/download.html). The pan-cancer ATAC-seq data set was obtained from TCGA GDC (https://gdc.cancer.gov/about-data/publications/ATACseq-AWG).

Code availability

Extensive documentation and a full user manual are available at https://github.com/haojiechen94/EAP/tree/main/doc. The software is open source, and all code can be found on GitHub at https://github.com/haojiechen94/EAP/tree/main/source_codes.

Author’s contributions

Z.S., Y.L., G.Z. and S.T. conceived the idea and designed the analysis platform. H.C., T.H., Z.G. and A.Z. developed the web server and analyzed the real datasets. H.C., Z.G., Z.S. and G.Z. wrote the manuscript. H.C. wrote the documentation with help from Z.G., W.C. and L.M.. All authors read and approved the final manuscript.

Funding

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (No. XDB38050200 and XDB38040100), the National Basic Research Program of China (2018YFA0800203) and the National Natural Science Foundation of China (32370705 and 32200519).

References

1 Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**, doi:10.1126/science.aav1898 (2018).

2 Wu, J. *et al.* The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature* **534**, 652-657, doi:10.1038/nature18606 (2016).

3 Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**, 669-680, doi:10.1038/nrg2641 (2009).

4 Grandi, F. C., Modi, H., Kampman, L. & Corces, M. R. Chromatin accessibility profiling by ATAC-seq. *Nat Protoc* **17**, 1518-1552, doi:10.1038/s41596-022-00692-9 (2022).

5 Lomberk, G. *et al.* Distinct epigenetic landscapes underlie the pathobiology of pancreatic cancer subtypes. *Nat Commun* **9**, 1978, doi:10.1038/s41467-018-04383-6 (2018).

6 Yi, G. *et al.* Chromatin-Based Classification of Genetically Heterogeneous AMLs into Two Distinct Subtypes with Diverse Stemness Phenotypes. *Cell Rep* **26**, 1059-1069 e1056, doi:10.1016/j.celrep.2018.12.098 (2019).

7 Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res* **48**, D882-D889, doi:10.1093/nar/gkz1062 (2020).

8 Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* **41**, D991-995, doi:10.1093/nar/gks1193 (2013).

9 Members, C.-N. & Partners. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2022. *Nucleic Acids Res* **50**, D27-D38, doi:10.1093/nar/gkab951 (2022).

10 Tu, S. *et al.* MAnorm2 for quantitatively comparing groups of ChIP-seq samples. *Genome Res* **31**, 131-145, doi:10.1101/gr.262675.120 (2021).

11 Sun, H. *et al.* Quantitative integration of epigenomic variation and transcription factor binding using MAmotif toolkit identifies an important role of IRF2 as transcription activator at gene promoters. *Cell Discov* **4**, 38, doi:10.1038/s41421-018-0045-y (2018).

12 Chen, H. *et al.* HyperChIP: identification of hypervariable signals across ChIP-seq or ATAC-seq samples. *Genome Biol* **23**, 62, doi:10.1186/s13059-022-02627-9 (2022).

13 Langmead, B. & Nellore, A. Cloud computing for genomic data analysis and collaboration. *Nat Rev Genet* **19**, 325, doi:10.1038/nrg.2018.8 (2018).

14 Boettiger, C. An introduction to Docker for reproducible research. *ACM SIGOPS Operating Systems Review* **49**, 71-79, doi:10.1145/2723872.2723882.

15 Lu, X. *et al.* Cell-lineage controlled epigenetic regulation in glioblastoma stem cells determines functionally distinct subgroups and predicts patient survival. *Nat Commun* **13**, 2236, doi:10.1038/s41467-022-29912-2 (2022).

16 Yuan, C. *et al.* A systematic dissection of the epigenomic heterogeneity of lung adenocarcinoma reveals two different subclasses with distinct prognosis and core regulatory networks. *Genome Biol* **22**, 156, doi:10.1186/s13059-021-02376-1 (2021).

17 Shi, X. *et al.* Integrated profiling of human pancreatic cancer organoids reveals chromatin accessibility features associated with drug sensitivity. *Nat Commun* **13**, 2169, doi:10.1038/s41467-022-29857-6 (2022).

18 Sanghi, A. *et al.* Chromatin accessibility associates with protein-RNA correlation in human cancer. *Nat Commun* **12**, 5732, doi:10.1038/s41467-021-25872-1 (2021).

19 Wang, Z. *et al.* The Open Chromatin Landscape of Non-Small Cell Lung Carcinoma. *Cancer Res* **79**, 4840-4854, doi:10.1158/0008-5472.CAN-18-3663 (2019).

20 Chen, T. W. *et al.* ChIPseek, a web-based analysis tool for ChIP data. *BMC Genomics* **15**, 539, doi:10.1186/1471-2164-15-539 (2014).

21 Smith, J. P. *et al.* PEPATAC: an optimized pipeline for ATAC-seq data analysis with serial alignments. *NAR Genom Bioinform* **3**, lqab101, doi:10.1093/nargab/lqab101 (2021).

22 Liu, T. *et al.* Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol* **12**, R83, doi:10.1186/gb-2011-12-8-r83 (2011).

23 Jiang, H., Wang, F., Dyer, N. P. & Wong, W. H. CisGenome Browser: a flexible tool for genomic data visualization. *Bioinformatics* **26**, 1781-1782, doi:10.1093/bioinformatics/btq286 (2010).

24 Fernandez, L. P., Lopez-Marquez, A. & Santisteban, P. Thyroid transcription factors in development, differentiation and disease. *Nat Rev Endocrinol* **11**, 29-42, doi:10.1038/nrendo.2014.186 (2015).

25 Garces de Los Fayos Alonso, I. *et al.* The Role of Activator Protein-1 (AP-1) Family Members in CD30-Positive Lymphomas. *Cancers (Basel)* **10**, doi:10.3390/cancers10040093 (2018).

26 Frigo, D. E. *et al.* Mechanism of AP-1-mediated gene expression by select organochlorines through the p38 MAPK pathway. *Carcinogenesis* **25**, 249-261, doi:10.1093/carcin/bgh009 (2004).

27 Fahiminiya, S., de Kock, L. & Foulkes, W. D. Biologic and Clinical Perspectives on Thyroid Cancer. *N Engl J Med* **375**, 2306-2307, doi:10.1056/NEJMc1613118 (2016).

28 Niu, D. F. *et al.* Transcription factor Runx2 is a regulator of epithelial-mesenchymal transition and invasion in thyroid carcinomas. *Lab Invest* **92**, 1181-1190, doi:10.1038/labinvest.2012.84 (2012).

29 Bailey, P. *et al.* Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* **531**, 47-52, doi:10.1038/nature16965 (2016).

30 Tiriac, H. *et al.* Organoid Profiling Identifies Common Responders to Chemotherapy in Pancreatic Cancer. *Cancer Discov* **8**, 1112-1129, doi:10.1158/2159-8290.CD-18-0349 (2018).

31 Erratum to: Impact of classical and basal-like molecular subtypes on overall survival in

resected pancreatic cancer in the SPACIOUS-2 multicentre study. *Br J Surg* **110**, 289, doi:10.1093/bjs/znac429 (2023).

32 Brunton, H. *et al.* HNF4A and GATA6 Loss Reveals Therapeutically Actionable Subtypes in Pancreatic Cancer. *Cell Rep* **31**, 107625, doi:10.1016/j.celrep.2020.107625 (2020).

33 Kalisz, M. *et al.* HNF1A recruits KDM6A to activate differentiated acinar cell programs that suppress pancreatic cancer. *EMBO J* **39**, e102808, doi:10.15252/embj.2019102808 (2020).

34 O'Kane, G. M. *et al.* Correction: GATA6 Expression Distinguishes Classical and Basal-like Subtypes in Advanced Pancreatic Cancer. *Clin Cancer Res* **28**, 2715, doi:10.1158/1078-0432.CCR-22-1272 (2022).

35 Wang, C. *et al.* Metformin inhibits pancreatic cancer metastasis caused by SMAD4 deficiency and consequent HNF4G upregulation. *Protein Cell* **12**, 128-144, doi:10.1007/s13238-020-00760-4 (2021).

36 Ozaki, T. *et al.* Impact of RUNX2 on drug-resistant human pancreatic cancer cells with p53 mutations. *BMC Cancer* **18**, 309, doi:10.1186/s12885-018-4217-9 (2018).

37 Cheng, Y. *et al.* RUNX1 promote invasiveness in pancreatic ductal adenocarcinoma through regulating miR-93. *Oncotarget* **8**, 99567-99579, doi:10.18632/oncotarget.20433 (2017).

38 Kawasaki, K. *et al.* An Organoid Biobank of Neuroendocrine Neoplasms Enables Genotype-Phenotype Mapping. *Cell* **183**, 1420-1435 e1421, doi:10.1016/j.cell.2020.10.023 (2020).

39 Masui, T. *et al.* Replacement of Rbpj with Rbpjl in the PTF1 complex controls the final maturation of pancreatic acinar cells. *Gastroenterology* **139**, 270-280, doi:10.1053/j.gastro.2010.04.003 (2010).

Figures

Figure 1 Overview of EAP architecture and the various analysis modules

(a) Inputs to EAP include raw sequencing data (in FASTQ format) and metadata (describes study design and sample phenotypes, in CSV format). EAP consists of Basic Analysis module and Advanced Analysis module. Basic Analysis module performs quality control, read mapping, peaks calling, creates a summary report for filtering poor quality samples and generates analysis-ready count tables, which are then used as inputs for Advanced Analysis module. A comprehensive collection of ChIP/ATAC-seq data analysis tools are encapsulated in the Advanced Analysis module. Each analysis tools will produce publication-ready results (figures and tables). (**b)** Workflow of Basic Analysis module, requires the upload of raw sequencing data and metadata using the Cloud Gene-Client tool. Upon completion of this analysis, analysis-ready count tables and a summary quality control report are generated for downstream analysis. **(c-d)** Two distinct research scenarios are illustrated: (c) For labeled data sets, differential analysis can be employed to identify differential signals among samples with different labels. Subsequently, differential TF motif enrichment analysis and differential TF activity analysis can be conducted to investigate the TFs linked to differential binding or open chromatin sites. (d) In datasets lacking pre-defined sample labels, hypervariable analysis can be utilized to identify hypervariable ChIP/ATAC-seq signals across the samples. These signals can then be leveraged for clustering analysis and cluster samples into distinct groups, and signature genes scoring analysis is utilized to characterize these clusters based on established gene sets. Supervised analysis tools can also be utilized to identify binding/open chromatin sites or transcriptional regulators specific to each cluster. **(e)** EAP offers a Data Set Browser that provides an interactive interface for convenient visualization of TF activity scores in each dataset.

Figure 2 Illustration of application of EAP on labeled data sets

(a) MA plots highlight the differentially accessible sites when comparing ATAC-seq profiles across normal tissues, tumor tissues, and metastatic tissues. **(b)** Volcano plots show the differential TF motif enrichment analysis results of differentially accessible sites between normal tissues and tumor/metastatic tissues. **(c)** Boxplot shows the MAPK pathway activities across different tissue types. **(d)** Two dimensional t-SNE visualization of ATAC-seq samples using hypervariable regions identified by EAP, samples are colored by tissue of origin. **(e-g)** Two dimensional t-SNE visualization of different ATAC-seq samples colored by TF activity scores calculated by EAP.

Figure 3 Illustration of application of EAP on label-free data sets

(a) Heatmap and dendrogram showing clustering of samples with similar chromatin accessibility profiles and clustering of peak regions with similar expression patterns. **(b-c)** Boxplots showing the expression activity of signature gene sets curated by previous studies. **(d)** Kaplan-Meier curves displaying differences in progression-free survival (PFS) between patients whose organoids defined as Basal-like or Non-basal-like based on ATAC-seq profiles. TLSs&LAs signatures score. **(e)** Heatmap of TF activity scores of top-ranked cluster specific TFs. **(f)** Boxplots showing the TF activity scores across samples in two public cancer related data sets, samples are grouped by tissue of origins (left panel) or subtypes defined EAP (right panel).

Supplementary figures

Supplementary figure 1 The login page and data transfer tools of EAP

(a) The login page of EAP (https://www.biosino.org/epigenetics/#/user/login), a demonstration account is available by click on "demo" in this page. **(b)** The data transfer client for utilization of EAP, which supports break-point resumable transfers. **(c)** A md5sum checking procedure has been integrated to ensure the integrity of uploaded files

Supplementary figure 2 Advanced Analysis module in EAP

User can choose the downstream analysis tools implemented in EAP and customize the configure parameters. The input and output examples demonstrate the comprehensive visualization and analysis results.

Supplementary figure 3 Illustrating the visualization of the TF activity scores of RUNX1/2/3 in two case study ATAC-seq data sets using the Data Set Browser

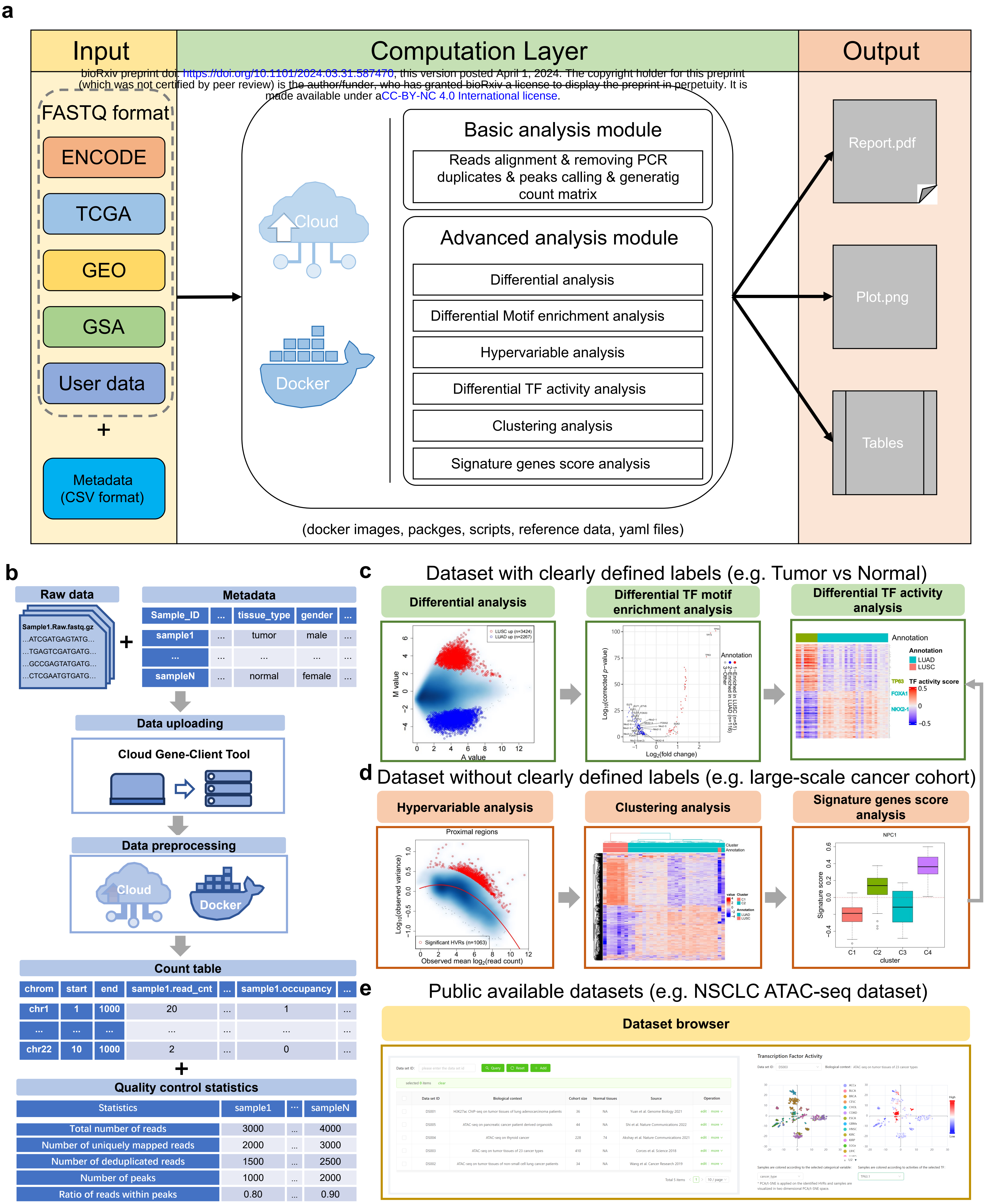
Supplementary movies

Supplementary_movie_1_Data_transferring_and_integrity_checking.mp4

Supplementary_movie_2_Basic_Analysis_module_demo_using_paired_end_ChIP_seq_data.mp4

Supplementary_movie_3_Advanced_Analysis_module_demo_differential_analysis.mp4

Figure 1



Metadata

Sample_ID	...	tissue_type	gender	...
sample1	...	tumor	male	...
...	...	...	...	...
sampleN	...	normal	female	...

Count table

chrom	start	end	sample1.read_cnt	...	sample1.occupancy	...
chr1	1	1000	20	...	1	...
...	...	...	...	...	...	...
chr22	10	1000	2	...	0	...

Quality control statistics

Statistics	sample1	...	sampleN
Total number of reads	3000	...	4000
Number of uniquely mapped reads	2000	...	3000
Number of deduplicated reads	1500	...	2500
Number of peaks	1000	...	2000
Ratio of reads within peaks	0.80	...	0.90

	Data set ID	Biological context	Cohort size	Normal tissues	Source	Operation
□	DS001	H3K27ac ChIP-seq on tumor tissues of lung adenocarcinoma patients	36	NA	Yuan et al. Genome Biology 2021	edit | more
□	DS005	ATAC-seq on pancreatic cancer patient derived organoids	44	NA	Shi et al. Nature Communications 2022	edit | more
□	DS004	ATAC-seq on thyroid cancer	228	74	Akshay et al. Nature Communications 2021	edit | more
□	DS003	ATAC-seq on tumor tissues of 23 cancer types	410	NA	Corces et al. Science 2018	edit | more
□	DS002	ATAC-seq on tumor tissues of non-small cell lung cancer patients	34	NA	Wang et al. Cancer Research 2019	edit | more

Data set ID: DS003 Biological context: ATAC-seq on tumor tissues of 23 cancer types

Samples are colored according to the selected categorical variable: cancer_type

* PCA/t-SNE is applied on the identified HVRs and samples are visualized in two dimensional PCA/t-SNE space.

Samples are colored according to activities of the selected TF: TP63.1

Figure 1 Overview of EAP architecture and the various analysis modules

(a) Inputs to EAP include raw sequencing data (in FASTQ format) and metadata (describes study design and sample phenotypes, in CSV format). EAP consists of Basic Analysis module and Advanced Analysis module. Basic Analysis module performs quality control, read mapping, peaks calling, creates a summary report for filtering poor quality samples and generates analysis-ready count tables, which are then used as inputs for Advanced Analysis module. A comprehensive collection of ChIP/ATAC-seq data analysis tools are encapsulated in the Advanced Analysis module. Each analysis tools will produce publication-ready results (figures and tables). (**b)** Workflow of Basic Analysis module, requires the upload of raw sequencing data and metadata using the Cloud Gene-Client tool. Upon completion of this analysis, analysis-ready count tables and a summary quality control report are generated for downstream analysis. **(c-d)** Two distinct research scenarios are illustrated: (c) For labeled data sets, differential analysis can be employed to identify differential signals among samples with different labels. Subsequently, differential TF motif enrichment analysis and differential TF activity analysis can be conducted to investigate the TFs linked to differential binding or open chromatin sites. (d) In datasets lacking pre-defined sample labels, hypervariable analysis can be utilized to identify hypervariable ChIP/ATAC-seq signals across the samples. These signals can then be leveraged for clustering analysis and cluster samples into distinct groups, and signature genes scoring analysis is utilized to characterize these clusters based on established gene sets. Supervised analysis tools can also be utilized to identify binding/open chromatin sites or transcriptional regulators specific to each cluster. **(e)** EAP offers a Data Set Browser that provides an interactive interface for convenient visualization of TF activity scores in each dataset.

Figure 2

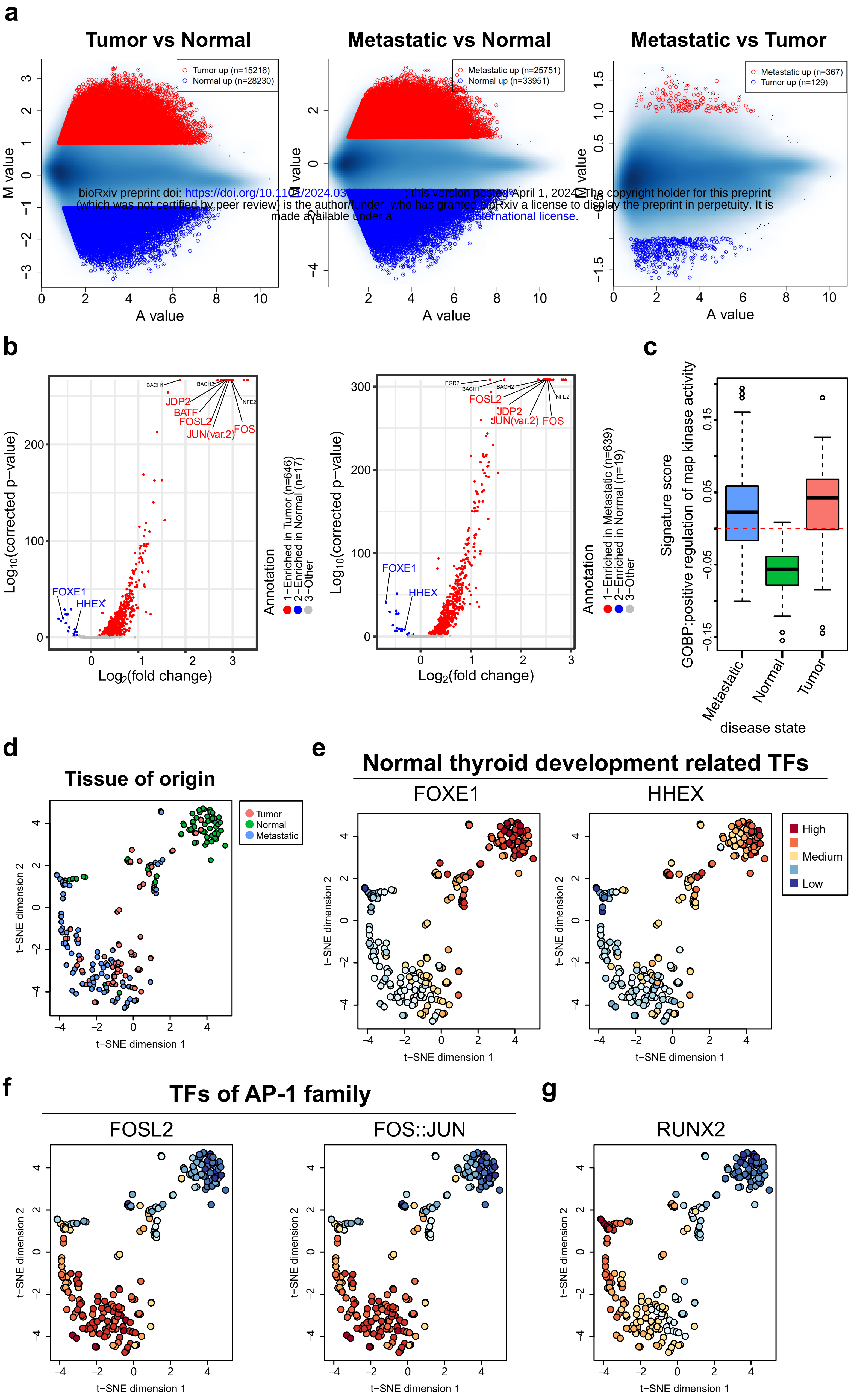

a
Tumor vs Normal
Tumor up (n=15216)
Normal up (n=28230)
M value
A value
Metastatic vs Normal
Metastatic up (n=25751)
Normal up (n=33951)
Metastatic vs Tumor
Metastatic up (n=367)
Tumor up (n=129)
b
Log10(corrected p−value)
Log2(fold change)
BACH1
BACH2
NFE2
JDP2
BATF
FOSL2
JUN(var.2)
FOS
FOXE1
HHEX
Annotation
1−Enriched in Tumor (n=646)
2−Enriched in Normal (n=17)
3−Other
EGR2
1−Enriched in Metastatic (n=639)
2−Enriched in Normal (n=19)
c
Signature score
GOBP:positive regulation of map kinase activity
Metastatic
Normal
Tumor
disease state
d
Tissue of origin
Tumor
Normal
Metastatic
t−SNE dimension 1
t−SNE dimension 2
e
Normal thyroid development related TFs
FOXE1
HHEX
High
Medium
Low
f
TFs of AP-1 family
FOSL2
FOS::JUN
g
RUNX2

Figure 2 Illustration of application of EAP on labeled data sets

(a) MA plots highlight the differentially accessible sites when comparing ATAC-seq profiles across normal tissues, tumor tissues, and metastatic tissues. **(b)** Volcano plots show the differential TF motif enrichment analysis results of differentially accessible sites between normal tissues and tumor/metastatic tissues. **(c)** Boxplot shows the MAPK pathway activities across different tissue types. **(d)** Two dimensional t-SNE visualization of ATAC-seq samples using hypervariable regions identified by EAP, samples are colored by tissue of origin. **(e-g)** Two dimensional t-SNE visualization of different ATAC-seq samples colored by TF activity scores calculated by EAP.

Figure3

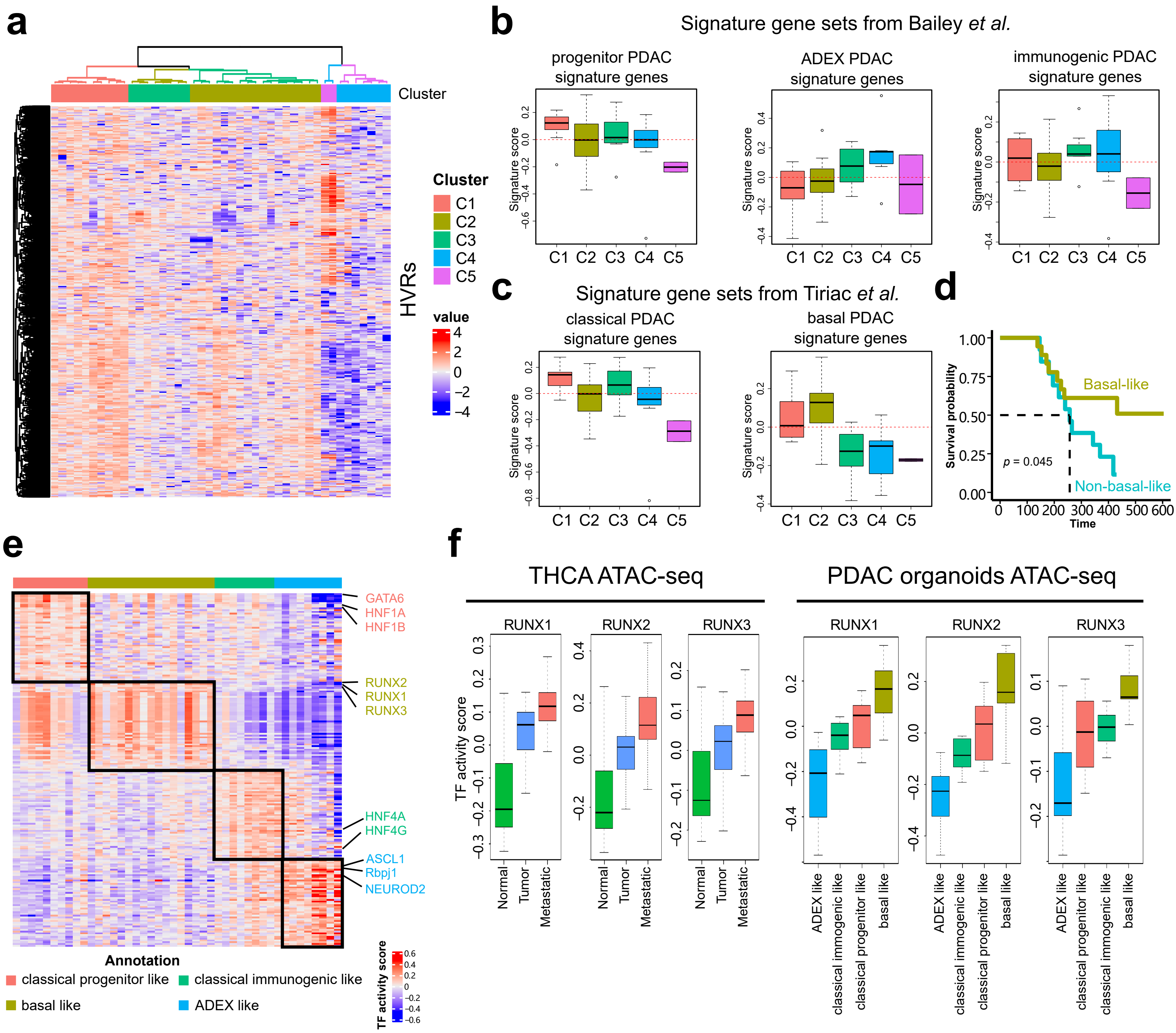

a
Cluster
HVRs
Cluster
C1
C2
C3
C4
C5
value
4
2
0
−2
−4
b
Signature gene sets from Bailey et al.
progenitor PDAC signature genes
ADEX PDAC signature genes
immunogenic PDAC signature genes
Signature score
C1 C2 C3 C4 C5
c
Signature gene sets from Tiriac et al.
classical PDAC signature genes
basal PDAC signature genes
Signature score
C1 C2 C3 C4 C5
d
Survival probability
1.00
0.75
0.50
0.25
0.00
Basal-like
Non-basal-like
p = 0.045
0 100 200 300 400 500 600
Time
e
GATA6
HNF1A
HNF1B
RUNX2
RUNX1
RUNX3
HNF4A
HNF4G
ASCL1
Rbpj1
NEUROD2
Annotation
classical progenitor like
classical immunogenic like
basal like
ADEX like
TF activity score
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
f
THCA ATAC-seq
PDAC organoids ATAC-seq
RUNX1
RUNX2
RUNX3
TF activity score
Normal
Tumor
Metastatic
ADEX like
classical immogenic like
classical progenitor like
basal like

Figure 3 Illustration of application of EAP on label-free data sets

(a) Heatmap and dendrogram showing clustering of samples with similar chromatin accessibility profiles and clustering of peak regions with similar expression patterns. **(b-c)** Boxplots showing the expression activity of signature gene sets curated by previous studies. **(d)** Kaplan-Meier curves displaying differences in progression-free survival (PFS) between patients whose organoids defined as Basal-like or Non-basal-like based on ATAC-seq profiles. TLSs&LAs signatures score. **(e)** Heatmap of TF activity scores of top-ranked cluster specific TFs. **(f)** Boxplots showing the TF activity scores across samples in two public cancer related data sets, samples are grouped by tissue of origins (left panel) or subtypes defined EAP (right panel).