Review

Integrative Epigenomics: Bioinformatics Strategies for Multi-Omics Data Analysis in Health and Disease

Shikhi Baruri *, Lalit Batra *, Sohome Adhikari, Ayman El-Baz

Posted Date: 7 May 2026

doi: 10.20944/preprints202605.0428.v1

Keywords: epigenomics; histone modification; DNA methylation; ChIP-seq; ATAC-seq; bioinformatics tools; epigenetic regulation; multi-omics integration; single-cell epigenomics; machine learning; clinical epigenetic biomarkers

 doi:10.20944/preprints202605.0428.v1

Review

Integrative Epigenomics: Bioinformatics Strategies for Multi-Omics Data Analysis in Health and Disease

Shikhi Baruri [1,*], Lalit Batra [1,*], Sohome Adhikari [2] and Ayman El-Baz [3]

[1] Center for Predictive Medicine for Biodefense and Emerging Infectious Diseases (CPM), University of Louisville, Louisville, KY 40222, USA
[2] Department of Computer Science and Engineering, University of Louisville, Louisville, KY 40292, USA
[3] Bioengineering Department, J.B. Speed School of Engineering, University of Louisville, Louisville, KY 40292, USA
* Correspondence: shikhi.baruri@louisville.edu (S.B.); lalit.batra@louisville.edu (L.B.)

Abstract

Background: Epigenomics has emerged as an essential technique of modern molecular biology, providing a critical layer of gene regulation. DNA methylation, histone modifications and alterations to the chromatin accessibility of DNA have been widely associated with complex diseases including cancer. The most recent developments in high-throughput sequencing technology have made it possible to profile epigenetic landscapes genomically on a large scale. However, these techniques often mask endogenous cellular heterogeneity essential for understanding complex disease stages. **Objective:** The purpose of the review is to assemble accessible tools and algorithms to conduct an Epigenomic study in the field of biomedical research, from bulk tissue analysis to the high-resolution frontier of single-cell epigenomics. **Methods:** We performed a comparative analysis of common methods used to analyze DNA methylation, chromatin immunoprecipitation, sequencing analysis and chromatin accessibility profiling. We described the standardized bioinformatics tools and pipelines required to transform raw sequencing data into mechanistic biological understanding, highlighting the role of quality control, peak calling, and differential analysis. Furthermore, we explore the integration of epigenomic with other "omics" layers through advanced computational frameworks, including machine learning and network-based modeling. Results: These advanced multi-omics techniques demonstrate efficient clinical utility by enabling biomarker discovery, disease subtyping, and the identification of novel therapeutic targets. **Conclusions:** Despite challenges with data complexity, the fusion of Artificial Intelligence (AI) and single-cell technologies will speed up the transition toward precision medicine. This review serves as a blueprint for providing the technical and computational complexities of epigenomics to uncover the mechanisms controlling human health and disease.

Keywords: epigenomics; histone modification; DNA methylation; ChIP-seq; ATAC-seq; bioinformatics tools; epigenetic regulation; multi-omics integration; single-cell epigenomics; machine learning; clinical epigenetic biomarkers

1. Introduction

Epigenomics is the study of the whole genome focusing on DNA methylation, histone modifications, chromatin structure, and noncoding RNAs. These mechanisms modulate gene expression without altering the underlying DNA sequence [1,2]. These epigenetic mechanisms are hereditary and reversible, and they are essential for regulating cellular identity, development, and tissue-specific functions by regulating the switches that turn genes on or off under different cellular contexts [3,4]. Consequently, abnormalities in epigenetic marks can cause complex diseases, such as cancer, cardiovascular, metabolic diseases, and immune-related diseases [1,5–8].

 doi:10.20944/preprints202605.0428.v1

More recently, next-generation sequencing (NGS) has accelerated epigenomic discovery by enabling genome-scale profiling of DNA methylation, histone modifications, and chromatin accessibility[9,10]. These advancements have facilitated the identification of epigenetic biomarkers for precision medicine. In fields such as oncology and cardiometabolic research, targeted epigenetic drugs are currently being explored to reverse disease-associated alterations. Furthermore, the emergence of single-cell epigenomics has provided a powerful method to decipher cellular heterogeneity and plasticity, offering critical insights into stem cell biology and tumor heterogeneity [3,4].

Large-scale epigenomic data, accessible through public repositories such as the NCBI Epigenomics database, allow researchers to explore epigenetics across diverse species and biological conditions [9]. By integrating genetic and environmental influences, epigenomics offers transformative potential for understanding complex biological systems, disease prevention, and personalized medicine [2,6,10].

However, the rapid expansion of epigenomic platforms has created an urgent need to evaluate the performance of computational tools across varied data types and analytical objectives. Specifically, the challenges regarding data quality, normalization, peak detection, functional annotation, batch effects, single-cell data sparsity and cross-platform integration still impede reproducibility and translational applications. These obstacles are particularly pronounced in multi-omics research, where epigenomic data must be integrated with transcriptomic, genomic, proteomic, or clinical datasets.

This review provides a critical examination of current epigenetic analysis tools, encompassing experimental platforms, computational pipelines, and commercially available technologies. We further explore machine-learning strategies including matrix factorization, network-based fusion, and deep learning, designed to enhance pattern recognition, biomarker discovery, and disease prediction. By reducing dimensionality and learning cross-modal relationships, these techniques allow researchers to extract meaningful biological insights from high-throughput assays. We also explore the role of epigenomics in cancer and other complex diseases, with particular emphasis on how epigenomic profiling is being implemented into early detection, diagnosis and therapeutic decision making. Lastly, rather than merely summarizing available tools, this review critically evaluates their applications, drawbacks and suitability for different epigenomic and multi-omics study designs. We also highlight new trends and future directions in the domain along with a clear view of how epigenomic research is defining the next generation of precision medicine.

2. Epigenomic Data and Analysis Pipelines

Epigenomic data is typically stored in formats that can be handled by genomic coordinates to enable integrative and comparative analysis. After processing, the information is represented as region sets or tracks that are mapped to reference genomes. That information is encoded in standard formats such as BED, bigWig, or narrowPeak files that indicate genomic regions containing specific epigenetic marks or quantitative data [9,11].

To simplify the handling of complex epigenomic datasets, especially in single-cell epigenomics, software packages like EpiScanpy have been developed to process single-cell DNA methylation and chromatin accessibility data. These packages include tools for clustering, dimension reduction, and differential analysis specifically designed for epigenomic properties [12] . Furthermore, data integration initiatives, such as the VISION project, assemble various epigenomic marks and transcriptomic data to produce large-scale maps of regulatory landscapes. These maps are essential for annotating functional genomic elements[13]. Resources like NCBI Epigenomics databases have thousands of epigenomic datasets available in standard formats to enable intuitive browsing, visualization, and large-scale genomic analysis [9]. Additionally, programmatic interfaces such as the DeepBlue Epigenomic Data Server allow researchers to easily access, filter, and summarize data from large consortia, including ENCODE and Roadmap Epigenomics, thereby supporting automated and reproducible workflows [11].

 doi:10.20944/preprints202605.0428.v1

Although current experimental approaches to epigenomics include different assays such as DNA methylation profiling, ChIP-seq, and chromatin accessibility mapping, most studies follow a standardized computational workflow **(Figure 1)** to ensure consistency and reliability in data analysis. The workflow begins with a quality assessment of raw sequencing reads, usually in FASTQ format, to assess the quality and integrity of data produced by the high throughput platforms. The filtration of low-quality reads, removal of adapter sequences, and elimination of potential contaminants are essential quality control steps to prevent errors in downstream analyses [14,15].

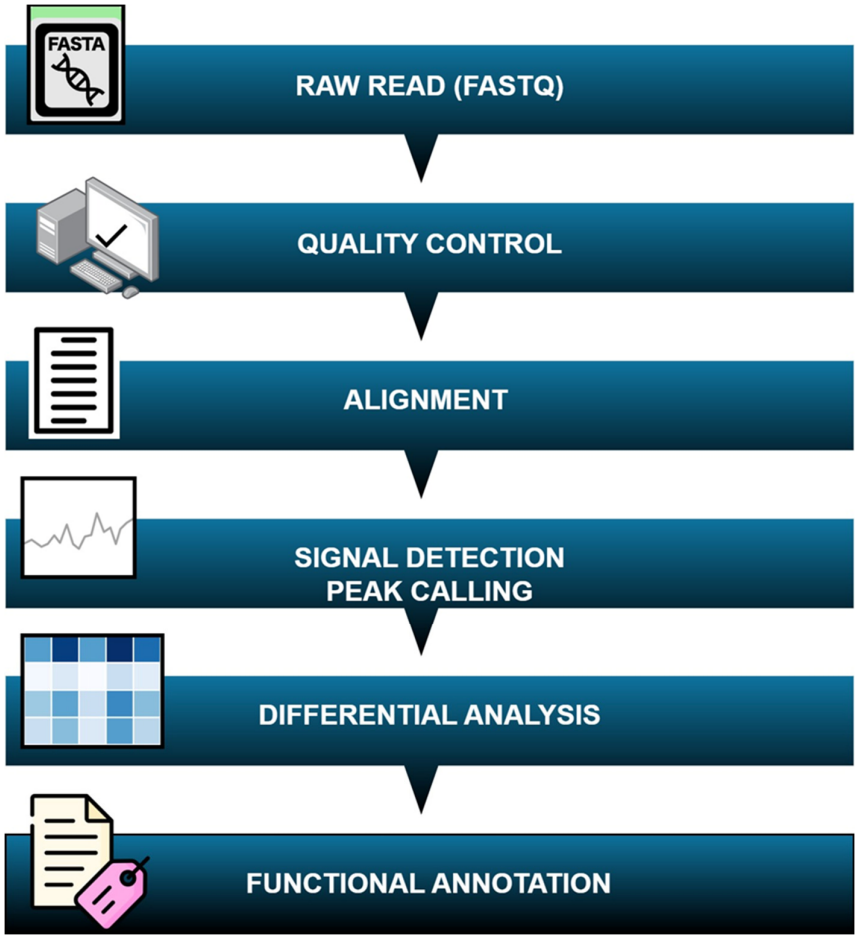


Figure 1. Standard Computational Pipeline for Epigenomic Data Processing. Schematic representation of the bioinformatic workflow used to transform raw sequencing output into biological insights. The process follows a modular hierarchy from raw reads to functional annotation like final biological interpretation.

Following quality control, the filtered reads are aligned to a reference genome using specialized alignment tools tailored to a specific assay. Accurate alignment is critical, because it determines the genomic locations from which the sequencing reads originate and forms the basis for signal detection. Once epigenetic signals are identified, differential analysis is performed to compare signals across different biological conditions (e.g., healthy versus diseased tissues) or experimental groups (e.g., treated versus control samples). Such comparisons can be used to identify regions of the genome that exhibit significant epigenetic changes, which are often associated with regulatory functions or disease mechanisms. The final stage of the workflow is the functional annotation, where identified epigenetic regions are linked to genomic features such as genes, promoters, enhancers, or other regulatory features. This outcome helps to decode how epigenetic changes influence gene expression and cellular activity.

 doi:10.20944/preprints202605.0428.v1

Every step of this workflow relies on specific bioinformatics tools chosen based on the assay type or research question. Errors or biases introduced in early stages, such as low-quality reads or misalignment, can propagate through the pipeline and lead to false biological findings. Therefore, pipeline-based decisions for epigenomic analysis are more than technical steps and are critical factors downstream inference that impact biological interpretation. For instance, normalization algorithms vary in how they handle sequencing depth and signal-to-noise ratios. Notably, the S3norm method was developed to normalize both parameters simultaneously and has been shown to capture biological variation more effectively than previous methods [13]. Similarly, interpretations of the same datasets can vary depending on the functional annotation strategy used, as reproducibility often differs across cell types [16]. These challenges are further intensified in single-cell epigenomics, where the feature space construction and integration strategies significantly influence clustering, cell-type discrimination, and downstream inference [12].

Standardization and reproducibility are still critical issues in epigenomic analysis as technical and biological variation can significantly impact data quality and interpretation. The effective sequencing depth and signal representation can be altered by differences in preprocessing parameters, such as read filtering, duplicate removal, and adapter trimming. The sensitivity and specificity of identified features are also affected by parameters such as reference genome selection, sequencing depth, peak-calling algorithm and threshold settings. In ChIP-based experiments, antibody quality represents a source of variability that cannot be addressed by computation. However, other technical factors such as batch effects in sample manipulation, library preparation, and sequencing conditions can further obscure the underlying biology. This issue is particularly acute in single-cell epigenomics, where low detection rates and sparse data matrices make results highly dependent on chosen strategies for feature construction, normalization, and integration [17,18].

The complexity is further compounded by cross-platform comparisons, which involve disparate regulatory features, signal types, resolutions, and data distributions. These variations stem from diverse measurement methods, including WGBS, ChIP-seq, ATAC-seq, CUT&RUN, CUT&Tag, and various single-cell assays. For example, WGBS measures base-level methylation, ChIP-based assays identify enrichment peaks for histone marks or DNA-binding proteins, ATAC-seq measures chromatin accessibility, and single-cell assays produce sparse and low-coverage matrices, that require modality-specific normalization and feature extraction. Because broad histone peaks, narrow transcription factor binding sites, open chromatin regions, and differentially methylated regions are not directly comparable units of analysis, direct cross-assay comparisons remain challenging.

Despite the ability of integrative frameworks like compEpiTools, methylPipe, and EpiScanpy to enhance harmonization by establishing shared feature spaces and annotation strategies, meaningful cross-platform integration still relies on deliberate normalization, batch correction, consensus reference annotations, transparent parameter reporting. Ideally, future studies will utilize standardized multi-assay profiling of identical biological samples to bridge these gaps [12]. Collectively, these problems demonstrate that epigenomic pipelines cannot be completely interchangeable; analytical decisions at every stage affect the reproducibility, interpretability, and comparability of studies. Increased standardization, clear reporting of workflow parameters, and standard benchmarking across assays and platforms are essential to facilitate sound biological inference and trustworthy clinical translation [18,19].

Practically, the choice of the workflow should be dictated by both the specific assay and the goal of analysis. Broad histone marks and narrow transcription factor binding events typically necessitate alternative peak-calling strategies, and single-cell assays necessitate workflows that are more robust to sparsity, low coverage, and modality-dependent noise. Similarly, normalization and annotation strategies optimized for bulk sequencing data may not be compatible with newer platforms, highlighting the necessity of assay-specific benchmarking[12,18,20].

Thus, the designing and implementation of standardized epigenomic pipelines **(Table 1)** are essential to improve inter-study reproducibility and providing robust insights into epigenetic regulation and its implications for health and disease [21–24].

Table 1. Epigenomics Tools Organized by Analysis Workflow. Example of graphical user interfaces tools for genomic tasks with applications, input/output format, advantages/ limitations. The table tracks the progression from raw sequencing data to functional biological insights across various high-throughput assays.

Step	Tool	Epigenetic Assay	Primary Application	Input	Output	Strengths	Limitations
Quality Control	FastQC	All epigenomic assays	Sequencing quality assessment	FASTQ	QC reports	Fast, user-friendly, widely accepted	Diagnostic only; no correction
	MultiQC	All assays	QC report aggregation	QC outputs	Combined QC report	Ideal for multi-sample studies	Depends on upstream QC tools
Alignment / Mapping	Bismark	Bisulfite-seq	DNA methylation alignment & calling	FASTQ	BAM, CpG methylation calls	High accuracy; gold standard	Computationally intensive
	BS-Seeker2	Bisulfite-seq	Methylation mapping	FASTQ	BAM, methylation calls	Faster than Bismark in some cases	Less commonly used
Signal Detection	MACS2	ChIP-seq, ATAC-seq	Peak calling	BAM	Peak (BED) files	Excellent background modeling; fast	Requires careful parameter tuning
	SICER	ChIP-seq	Broad peak detection	BAM	Broad peak regions	Well suited for histone marks	Less sensitive for narrow peaks
Differential Analysis	methylKit	Bisulfite-seq	Differential methylation analysis	Methylation calls	DMRs, plots	Robust statistics; flexible designs	Sensitive to low coverage
	DiffBind	ChIP-seq	Differential binding analysis	Peak files	Differential peaks	Integrates DESeq2/edgeR	High memory usage
Motif & Accessibility QC	HOMER	ChIP/ATAC-seq	Motif discovery	Peak files	Motifs	Strong motif analysis	Steep learning curve
	ATACseqQC	ATAC-seq	ATAC-seq quality assessment	BAM	QC metrics	Tailored for ATAC-seq	R-based complexity

Single-Cell Analysis	ArchR	scATAC-seq	Single cell epigenomics	Fragment files	Clusters, peaks	Powerful and scalable	High computational demand
	Signac	scATAC-seq	scATAC + scRNA integration	scATAC data	Integrated objects	Excellent multi-omics integration	Requires Seurat expertise
	SnapATAC	scATAC-seq	Single-cell accessibility	FASTQ/BAM	Cell clusters	Handles large datasets	Complex workflow
Genomic Operations	BEDTools	All assays	Genomic region operations	BED/BAM	Intersections	Flexible and fast	Command-line expertise needed
Functional Annotation	ChIPseeker	ChIP-seq	Peak annotation	BED peaks	Gene annotations	Easy gene association	Annotation dependent
	GREAT	ChIP/ATAC-seq	Regulatory region interpretation	Peaks	Functional enrichment	Regulatory-focused analysis	Web-based limitations
Multi-omics Integration	MOFA	Multi-omics	Data integration	Multi-omics matrices	Latent factors	Integrates diverse datasets	Interpretation complexity
	mixOmics	Multi-omics	Feature selection & integration	Multi-omics data	Integrated models	Strong visualization	Requires statistical expertise

 doi:10.20944/preprints202605.0428.v1

3. Comparative Evaluation of Epigenomics Tools

Despite the availability of numerous wet lab epigenetic tools **(Figure 2)** for specific analytical stages, their performance remains dependent on the assay type, signal properties, and size of data to be analyzed. Therefore, a meaningful comparison must not only focus on the major function of a tool but also consider its reproducibility, scalability, implementation, downstream compatibility, and applicability to a specific experimental design. No single tool is universally superior; rather, choices depend largely on requirements such as sensitivity, interpretability, efficiency, and ease of integration into existing workflows.

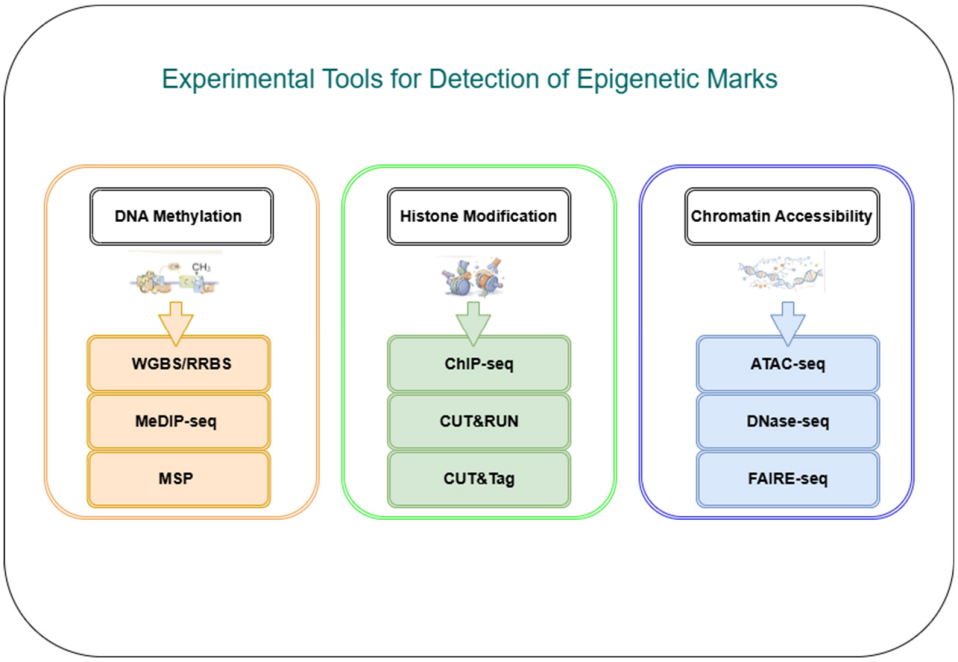


Figure 2. Experimental Modalities for Profiling Epigenetic Landscapes. A hierarchical classification of experimental methodologies used to detect and quantify various epigenetic modifications using wet lab. The figure categorizes tools into three primary biological domains.

3.1. Tools Based on Epigenetic Markers

For epigenomic analysis, two widely used tools for DNA methylation are Bismark, followed by MethylKit. WGBS (Whole Genome Bisulfite Sequencing) and RRBS (Reduced Representation Bisulfite Sequencing) DNA methylation profiling techniques can be used to determine CpG methylation patterns genome-wide [25–27]. Additionally, MeDIP-seq (Methylated DNA Immunoprecipitation Sequencing) identifies methylated DNA sites using 5-methylcytosine antibodies [25,28]. MSP (Methylation-Specific Polymerase Chain Reaction) remains one of the oldest techniques of identifying the presence of DNA methylation at precise locations [28].

Bismark is a methylation caller specifically designed for bisulfite sequencing that offers perform flexible alignment. It efficiently maps sequencing reads generated from bisulfite-treated DNA and performs methylation calling in a single step. This tool allows researchers to discriminate cytosine methylation across different sequence contexts such as CpG, CHG, and CHH and can provide detailed methylation profiles at single-base resolution. This type of functionality allows the visualization and interpretation of methylation throughout the genome. Furthermore, Bismark is open source, freely available, and supports time efficient processing, making it widely accessible for many laboratories [29].

However, Bismark has limitations, as it is heavily dependent on the quality of the data generated from bisulfite-treated DNA. Aligning large bisulfite sequences requires a significant amount of

 doi:10.20944/preprints202605.0428.v1

processing time and resources for whole-genome analyses. While Bismark is popular due to its alignment accuracy and identification capabilities, however it can be expensive to use for whole-genome datasets [29]. In comparison, BS-Seeker2 can be faster in certain environments, but it is less widely used in practice **(Table 2)**.

Table 2. Comparative analysis of widely used DNA methylation sequencing technologies. Comparison of common laboratory methods for profiling DNA methylation, categorized by their underlying biochemical principles, practical benefits, and technical constraints.

Method	Principle	Advantages	Limitations
WGBS	Bisulfite conversion and sequencing of almost the entire genome	Considered the gold standard for unbiased methylome profiling. Single-base resolution and near-comprehensive genome-wide coverage can be achieved.	Expensive and bisulfite treatment may degrade[138].
RRBS	Restriction enzyme enrichment of CpG-rich regions followed by bisulfite sequencing	More cost-effective than WGBS. High resolution in CpG islands and promoters. Lower sequencing requirement	Covers only a subset of the genome; biased toward CpG-dense regions; may miss intergenic/distal regulatory CpGs [138].
MSP	PCR amplification of bisulfite-converted DNA using primers specific for methylated sequences	Fast, inexpensive, highly sensitive and suitable for low input samples.	Limited to predefined loci. It is not genome wide. Also, it depends heavily on primer design. It may provide only qualitative or semi-quantitative information [139,140].

To perform downstream differential methylation analysis, methylKit offers a powerful statistical framework and experimental flexibility. However, it has limitations such as its performance can be affected by low depth and coverage sensitive regions. Consequently, selecting tools for methylation analysis often involves a trade-off between accuracy of alignment, runtime, adoption ease, statistical flexibility, and dataset size.

3.1.1. Chromatin and Histone Analysis

Histone modification mapping is typically performed by ChIP-seq and more recent enzyme-tethering methods like CUT&RUN and CUT&Tag [30–32]. Numerous computational tools exist for ChIP-seq data, focusing on data processing, peak calling, normalization, visualization, and downstream analysis. In case of ChIP-seq data analysis, tools like PeakSeq and DROMPA (Discrepancy Restricted Overlapping Mapped Peak Analyzer) are used for peak calling. These tools identify and score protein-binding regions to DNA, along with control samples. PeakSeq uses a 2-step method for sequence mapping to control open chromatin bias, which improves experimental design and sequencing depth estimation [33,34].

DROMPA is designed for efficient peak calling and visualization, particularly for identifying binding profiles in repetitive sequences and handling both broad and sharp peaks. Moreover,

 doi:10.20944/preprints202605.0428.v1

DROMPA is user friendly and accessible to users with limited bioinformatics background [35]. In chromatin profiling, signal shape and specific analysis objectives are critical factors in selecting the most appropriate tool. PeakSeq offers better signal detection by correcting background noise. Moreover, it is valuable when there is a concern about open-chromatin bias. DROMPA is highly efficient at identifying and visualizing peaks, making it especially useful for studying both broad and sharp peak architecture within a user-friendly framework.

Tools like CoBRA (Combined Bisulfite Restriction Analysis) offer organized bioinformatics workflows for analyzing ChIP-seq and ATAC-seq data. These tools streamline several critical steps, including normalization, correcting copy number variations, clustering, differential peak calling, identifying motifs, and analyzing pathways. By integrating these processes, researchers can better understand chromatin accessibility and the complexities of protein-DNA interactions [36].

Other resources, such as EaSeq offers interactive environments that combine genome browsing with user-friendly tools for data exploration, abstraction, and visualization. This transparency allows researchers to generate hypotheses and analyze large datasets in a reproducible manner [37]. Ultimately, a systematic evaluation of over 30 computational tools for differential ChIP-seq analysis indicates that tool performance depends significantly on the size and shape of peaks as well as specific biological situations. This emphasizes the need for tailored selection of algorithms based on the experimental design and biological question to maximize both accuracy and usability [34,38].

3.1.2. Assay-Specific and Advanced Methods

Advanced methods like MAPS (Model-based Analysis of PLAC seq) have been developed for integrated analyses combining ChIP with chromatin conformation approaches (e.g., PLAC-seq, HiChIP). By applying statistical modeling, MAPS identifies long-range chromatin interactions anchored at protein-bound sites, significantly expanding the insights gained from standard ChIP-based assays [39].

Recent advancements have introduced assay-specific tools that further enhance chromatin profiling workflows. SEACR (Sparse Enrichment Analysis for CUT&RUN) serves as a selective peak caller for CUT&RUN data [40] and GoPeaks is tailored for selective peak caller for CUT&Tag histone-modification profiling [41]. LanceOtron, a machine-learning-driven peak caller that is more advanced than traditional methods, has been used in recent benchmarking studies to compare the performance across diverse chromatin profiling datasets [42]. These more recent tools point to the ongoing move to assay-specific and learning-based methods of detecting peaks. For example, Omnipeak has just been released to call chromatin-mark peaks on both narrow and broad signal profiles, with a focus on peak stability and reproducibility. These findings highlight the significance of tools selection based on peak architecture and experimental design [43]. The choice of tool should consider factors such as data characteristics, biological questions, computational resources, and user expertise to effectively interpret protein-DNA interaction landscapes [34,36–39].

Comparative analyses of the differential ChIP-seq readings have revealed that no single method outperforms others across all peak architectures and biological contexts. Instead, tool selection should be meticulously tailored based on specific criteria such as shape width of the peaks, experimental design, computational resources, and downstream objectives [33,35,38].

3.1.3. ATAC-Seq and Regulatory Analysis

ATAC-seq is a vital technique for identifying regulatory regions, including enhancers and promoters. The rapid adoption of ATAC-seq has been accompanied by the development of specialized computational tools tailored to process, analyze, and interpret ATAC-seq data effectively. MACS2, originally developed for ChIP-seq peak calling, has been adapted for ATAC-seq to identify regions of accessible chromatin by detecting significant enrichments in sequencing reads [30,44]. It remains a robust and widely used peak caller for ATAC-seq data, particularly effective when enrichment profiles are relatively sharp [45,46]. HOMER (Hypergeometric Optimization of Motif EnRichment) is frequently applied after peak calling to perform motif discovery within accessible

 doi:10.20944/preprints202605.0428.v1

regions. It allows researchers to infer which transcription factors and regulatory elements may bind to open chromatin sites [46].

GenomicRanges is a software tool for region-based comparisons that helps analyze identified peaks in the context of genomic features, supports overlap queries, and assists in integrating ATAC-seq peaks with other genomic datasets. This enhances the ability to compare chromatin accessibility across different conditions or cell types [46].

FAIRE-seq is a specialized technique of determining open chromatin regions by isolating the nucleosomes-depleted DNA[47]. Complementary tools, such as ATAC graphs provide ATAC-seq-specific functionalities, including quality control, differential accessibility analysis, and peak identification of both nucleosome-free and nucleosome-occupied regions. Designed with user-friendly interfaces, these tools are particularly accessible to bench scientists who have minimal experience in bioinformatics [48].

HMMRATAC uses a Hidden Markov Model to find accessible regions within DNA. Unlike standard methods, it can accurately distinguish between nucleosome-free DNA and nucleosome-occupied DNA throughout the genome. This method is more sensitive than generic peak calling [49]. Comprehensive pipelines such as CoBRA support reproducible and scalable workflows for both ChIP-seq and ATAC-seq analysis. These pipelines integrate several critical analytical stages, including normalization, differential peak calling, motif enrichment, and pathway analysis to efficiently interpret chromatin accessibility profiles [36].

These tools allow researchers to map the regulatory framework such as active enhancers, promoters, and binding sites of transcription factors **(Table 3)** which further helps to understand gene regulation and cellular identity [46,50]. The selection of an ATAC-seq workflow is not universal; rather, it should be tailored to the specific biological question and the expected profile of the chromatin signal. MACS2 is an effective and most widely used general-purpose peak caller, especially for accessible chromatin regions with sharp enrichment profiles. However, HMMRATAC can be more beneficial when it is necessary to distinguish between nucleosome-free and nucleosome-bound regions. HOMER is then preferred choice for discovering downstream motifs, whereas GenomicRanges is useful when comparative and integrative analyses across conditions or datasets are required. Ultimately, ATAC-seq workflows are most effective when selected based on the specific requirements for peak detection sensitivity, chromatin-state resolution, motif analysis, or cross-dataset comparisons [45,46,49].

3.2. Tools for Functional Annotation and Data Interpretation

To interpret the biological significance of epigenomic signals, researchers employ functional annotation and integration tools that connect these signals to specific genomic features and regulatory functions. BEDTools is an effective and highly versatile program that can be used to either intersect or compare genomic features (epigenomic peaks), allowing for the annotation of regions in BED or GFF formats. This enables the determination of the overlap between the epigenetic marks and the known genomic elements [51]. ChIPseeker is an R/Bioconductor package that is specifically used to annotate ChIP-seq or other epigenomic peaks with nearest genes and genomic features such as promoters, exons, and introns. Beyond basic annotation, it offers advanced capabilities including peak comparison and visualization around the transcription start sites. By leveraging thousands of publicly available datasets, ChIPseeker assists in identifying co-regulation patterns and transcription factor complexes [52,53]. GREAT (Genomic Regions Enrichment of Annotations Tool) enhances the interpretation of cis-regulatory region functions by incorporating both the proximal and distal binding sites of genes [54]. Unlike tools that rely solely on “nearest-gene” relationships, GREAT employs sophisticated statistical models to assess enrichment across various ontologies. This provides researchers with deeper insights into the biological processes and pathways associated with specific epigenomic regions.

Table 3. Comparative Evaluation of Representative Epigenomics Analysis Tools. This table covers the primary stages of the analytical pipeline, including read alignment, methylation calling, peak identification, differential binding analysis, and functional annotation. Each tool is evaluated based on its technical strengths, inherent limitations, and the specific research context in which it performs most reliably.

Tool	Application	Strengths	Limitations	Optimal Use Case	Developer / Platform	Key reference(s)
Bismark	DNA methylation analysis	Accurate CpG methylation detection; supports both directional and non-directional libraries; well established in the epigenomics community	Computationally intensive; requires careful parameter optimization.	High-confidence bisulfite alignment and integrated methylation calling are priorities, especially for established WGBS/RRBS workflows	Babraham Bioinformatics	[29]
methylKit	Differentially methylated region (DMR) detection	Robust statistical framework; supports multiple experimental designs; integrates well with downstream visualization tools	Coverage sensitive; performance affected by low read depth; requires familiarity with R	Flexible differential methylation analysis is needed across multiple sample groups or experimental designs	Bioconductor	[141]

MACS2	Peak calling	**Widely adopted; effective background noise modeling; applicable to both narrow and broad peaks; fast and scalable**	**Performance depends on control input quality; parameter tuning required for different histone marks**	General-purpose peak calling is needed, especially for relatively sharp enrichment profiles or standard ChIP-seq/ATAC-seq workflows	Open-source (GitHub)	[142]
DiffBind	Differential chromatin binding analysis	**Flexible design; identifies differential binding events; integrates with DESeq2 and edgeR; supports complex experimental designs**	**Computationally intensive for large datasets; requires high-quality peak calls**	Differential binding analysis is the main goal and good-quality peak sets are already available	Bioconductor	[143]
ChIPseeker	Peak annotation	**Easy integration with downstream analysis; efficient peak-to-gene association**	**Annotation dependent**	Rapid gene-centered annotation and peak-distribution analysis are needed after peak calling	Bioconductor	[53]

 doi:10.20944/preprints202605.0428.v1

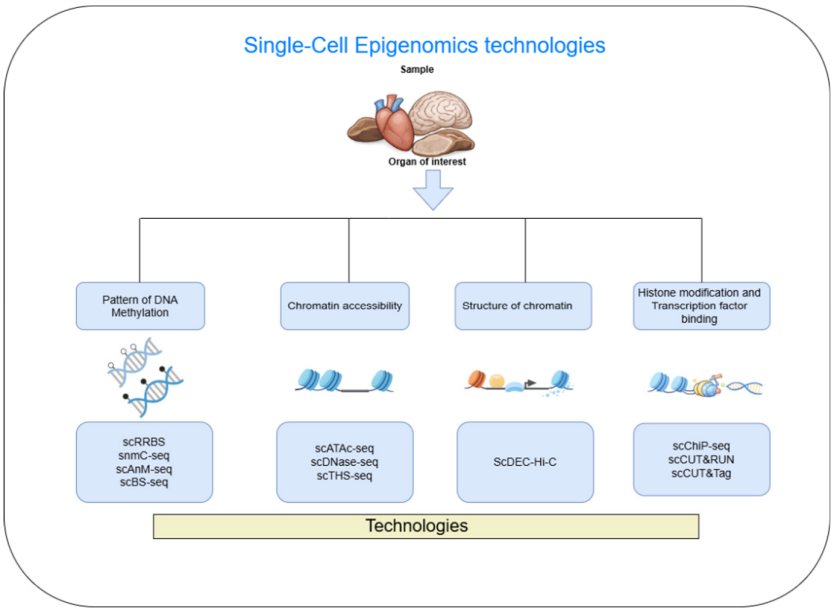
GREAT is available as a web application and designed to manage false positive results in mapping of genome-wide data sets and is useful not only to ChIP-seq but can also be applied to open chromatin and other data types of epigenomic data [54] . These tools allow extensive linking of epigenetic alterations to gene regulation and biologic pathways, allowing researchers to gain valuable functional information from raw epigenomic data [51,52,54]. Together with tools like BEDTools and ChIPseeker, GREAT enables researchers to extract valuable functional information from raw epigenomic data, providing a clearer picture of how the epigenome influences cellular behavior.

Functional annotation tools are required for interpreting epigenomic regions, yet their utility depends heavily on the specific annotation approach used While overlap-based methods like BEDTools are versatile and effective for intersecting peaks with known genomic features, they often fail to identify which specific genes or regulatory programs are biologically affected [51]. The nearest-gene annotation and peak distribution analysis is easily detected by ChIPseeker, but nearest-gene assignment can be ambiguous on distal enhancers, intergenic peaks, or loci where multiple potential targets exist [52,53]. GREAT partially overcomes this shortcoming by including both proximal and distal regulatory domains, although its interpretations remain dependent on pre-defined association rules and may yield false positive when regulatory regions are indirectly associated with genes [54].

These limitations are particularly significant because most epigenomic signals reside in non-coding regions. In these areas, functional implications are often context-dependent and hard to understand without integrating supplementary data such as chromatin conformation data, expression correlation, or multi-omics integration. Consequently, annotation results must be interpreted with caution, and tools should be selected based on the specific research objective—whether that is a rapid genomic overlap, gene-focused annotation, or a broad regulatory analysis.

3.3. Tools for Single-Cell Epigenomic Technologies

Recent advancements in sequencing technologies have moved epigenomic profiling beyond bulk analysis, which often masks critical cellular heterogeneity by averaging signals across large cell populations. Single-cell epigenomics provides a high-resolution view of regulatory dynamics, capturing cell-to-cell variations in DNA methylation, chromatin accessibility, histone modifications, and chromosomal conformations. These methods (**Figure 3)** are essential for understanding cellular identity, plasticity, and tissue-specific functions in complex biological systems.



 doi:10.20944/preprints202605.0428.v1

Figure 3. Illustration of single-cell epigenomics technologies. Schematic workflow for the application of single-cell sequencing technologies across various epigenetic layers. Tissues are dissociated to enable the profiling of individual cells, bypassing the averaging effect of bulk sequencing. These technologies are categorized into four primary functional domains.

3.3.1. Computational Frameworks for Profiling Epigenetic Features at Single-Cell Resolution

scATAC-seq and single-cell DNA methylation sequencing allow the identification of regulatory states in individual cells. ArchR, Signac and SnapATAC are advanced computational tools developed to support key analytical tasks in single-cell chromatin accessibility studies. These platforms support essential analytical tasks, including peak calling, dimensionality reduction, cell clustering, and integration with single-cell transcriptomic data [55–57]. A variety of specialized assays allow for the detailed mapping of individual epigenetic features such as DNA methylation, chromatin accessibility, histone modifications and chromosomal conformations for each cell. These methods help to understand cellular identity, plasticity and tissue-specific functions.

The pattern of DNA methylation in single cells can be identified by the scWGBS (single-cell whole genome bisulfite sequencing), single-cell reduced representation bisulfite sequencing (scRRBS), single-cell bisulfite sequencing (scBS-seq), and single-nucleus methylcytosine sequencing (snmC-seq). This capability permits the investigation of heterogeneity in DNA methylation on individual cell basis, providing deeper insight into cell-to-cell epigenetic variation. This approach facilitates a deeper understanding of dynamic changes in cell states during processes such as differentiation and in disease contexts, including cancer [58]. Chromatin accessibility can be performed with assays like single-cell ATAC-seq (scATAC-seq), scDNase-seq, and scTHS-seq, that define the nucleosomes-depleted regions and regulatory elements like promoters and enhancers. Histone modifications and transcription factor binding can be analyzed by using single-cell ChIP-seq, CUT&RUN and CUT&Tag, which enables the mapping of proteins and regulatory complexes associated with the chromatin in individual cells. ScDEC-Hi-C (single-cell deep embedded clustering) can be used to analyze the structure of chromatin and long-range genomic interactions in three-dimensional architecture by measuring spatial proximity between DNA loci [44,59–61].

3.3.2. Emerging Frontiers: Spatial and Genomic Amplification

For complex tissues, spatially resolved single-cell epigenomics have emerged to link gene regulation with physical location. For instance, in situ segmentation and multiplexed imaging enable the profiling of histone modifications and enhancer activity in the individual cells across tissue sections, thus linking spatial gene regulation with developmental and functional biology [48].

Furthermore, analyzing the limited DNA found in single cells is made possible through whole-genome amplification techniques like MALBAC (Multiple Annealing and Looping-Based Amplification Cycles) or LIANTI (Linear Amplification via Transposon Insertion). These methods ensure reliable genomic profiling even from extremely small quantities of starting material. Besides genomic information, a variety of single-cell epigenomic methods have been developed to profile individual epigenetic features.

3.3.3. Challenges in Single-Cell Analysis

However, single cell epigenomics data analysis is subject to several limitations, including primarily high technical noise, sparsity of data matrices, and increased computational complexity. Firstly, it is often challenging to differentiate true biological signals from artifacts due to high technical noise in single cell epigenomic data. Addressing this requires specialized noise models and robust computational frameworks to enhance data accuracy and reliability [62]. Secondly, data sparsity is a hallmark of single-cell epigenomics. For instance, in scATAC-seq data, around 1 to 10% of chromatin accessibility peaks are detected per cell due to limited sequencing depth and DNA sampling from diploid genomes. These highly sparse matrices often hinder essential downstream

 doi:10.20944/preprints202605.0428.v1

tasks such as cell-type identification and clustering [63]. Hence, special tools are required to account for such sparsity to recover meaningful regulatory features and to enhance data quality [63,64].

Finally, the high dimensionality and large scale of single cell epigenomic datasets eventually lead to increased computational complexity. Processing such large-scale, noisy data requires advanced algorithms, including deep learning, graph neural networks, and variational autoencoders. Moreover, they must reduce data size, correct batch errors, and interpret biological data efficiently. Despite recent advances, the complexity of these analyses remains a bottleneck, especially for the integration of multimodal single-cell omics datasets [62,65,66].

3.3.4. Specialized Single-Cell Analysis Tools

To address these limitations, several specialized tools have been developed such as, SCAVENGE (Single Cell Analysis of Variant Enrichment through Network propagation of GEnomic data) for mapping genetic variants to single-cell states, scCASE (single-cell Chromatin Accessibility Sequencing Enhancement) for data enhancement and mitigate sparsity, and casTLE (Cas9 High-Throughput Maximum Likelihood Estimator) for generating meaningful latent embedding to improve the precision of single-cell analysis [53,64,67] .

4. Integration of Epigenomics with Other Omics Data

Human health and disease are too complex biologically to be entirely represented by any single omics modality. The integration of epigenomics with genomics, transcriptomics, proteomics, radiomics and clinical data helps to understand more complex biological insights through the integration of multiple molecular data. The most common strategies of computational integration of heterogeneous datasets are statistical-based approaches (principal component analysis (PCA) or canonical correlation analysis (CCA), network-based approaches (such as gene regulatory and protein interactions networks), and machine learning algorithms (like clustering algorithms or deep learning models). These analytical tools facilitate the identification of disease-associated molecular signatures, thereby permitting biomarker discovery, disease subtyping, and identification of therapeutic targets **(Figure 4)**.

These integrative methods offer a system-level perspective of disease pathogenesis and serve as a foundation to build precision medicine strategies. Moreover, the integration of multi-omics data has, in recent years, greatly facilitated the development of precision medicine strategies [68]. The selection of a specific strategy is usually based on biological questions, like data type (e.g., bulk or single cell), and the required level of interpretability. Computational integration strategies are important in the analysis of multi-omics data as they enable the combined integration and interpretation of different high-dimensional molecular data sets.

Practically, the integration strategies differ not only in their mathematical foundations but also in their assumptions, interpretability, computational load, and specific applications. For exploratory analysis, PCA and CCA are typically used to perform dimensionality reduction of data and detecting shared variance across modalities. However, matrix factorization is preferred when interpretable latent factors or tolerance for missing data are required. Network-based approaches are typically the most suitable for patient stratification and relationship studies, whereas deep learning models are usually the most beneficial when the targeted complex nonlinear interactions are studied in large and high-dimensional multimodal data. Ultimately, method selection **(Table 4)** must balance predictive performance with scalability, robustness, and the ability to harmonize heterogeneous data [69–72].

4.1. Statistical Methods

PCA and CCA are fundamental tools, which allow the reduction of dimensions and determining the relationship between modalities. PCA removes noise and redundancy by converting the original features into major components and this enables further classification or regression frameworks. For instance, combination of nonnegative PCA with support vector machines (NPCA-SVM), [73]has

 doi:10.20944/preprints202605.0428.v1

proven robust in identifying cancer biomarkers in microarray data. CCA, determines the relationship between two or more types of omics data, can be combined with machine learning to extract canonical variables. These variables serve as an integrated and compact representation of shared biological signals across datasets [74,75]. Advanced variations, such as L2,1-norm constrained CCA, boosts the feature selection and noise resistance, which allows easier and more discriminative and compact data representation in machine learning applications [76]. Nonlinear association discovery and feature extraction of high dimensional datasets are made possible through kernel and sparse implementations of CCA. These extracted features can then be used as input for machine learning classifiers to capture nonlinear relationships among biomarkers, thereby improving patient stratification [77].

As an example, Supervised Deep Generalized Canonical Correlation Analysis (SDGCCA) has outperformed traditional CCA in classifying Alzheimer's disease and various cancers [75]. By improving signal extraction, robustness, and interpretability, machine learning models based on features extraction from PCA and CCA have facilitated clinically validated biomarker discoveries [78].

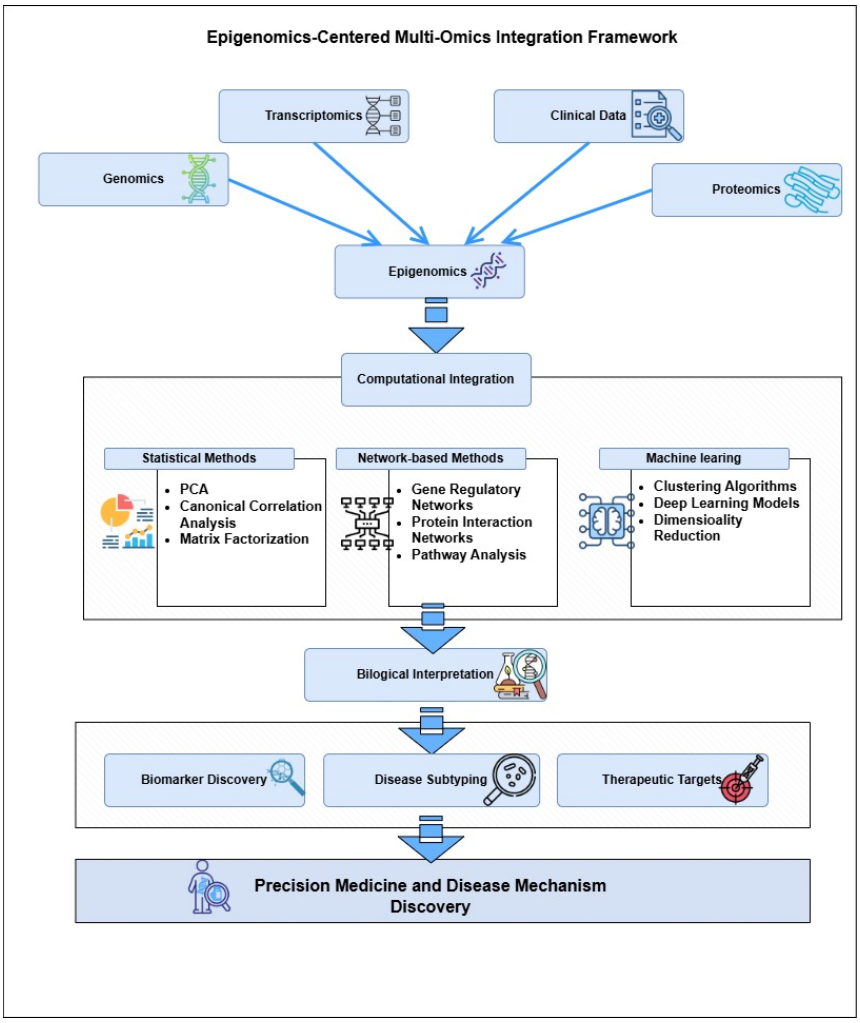


Figure 4. Representation of the Multi-Omics Integration Framework. Strategies enabling complex analysis of biological systems through the integration of multiple molecular datasets. A conceptual workflow demonstrates the convergence of diverse biological data streams through computational intelligence to drive clinical discovery. The process is divided into data acquisition, computational integration, biological interpretation and clinical applications.

Table 4. Summary of Integrative Omics Tools and Computational Strategies. This table provides a comparative overview of key computational tools used for the integration of epigenomics with other omics modalities. It details the underlying mathematical strategies, the types of datasets typically employed, key performance findings, and the specific research contexts where each tool is most effective.

Tool name	Integration strategy	Dataset used	Key findings	Best used when	Interpretability / trade-off
MOGONET	Graph convolutional networks / supervised deep learning	mRNA expression, DNA methylation, and miRNA expression across biomedical classification datasets	More effective than the other state of the art supervised multi omics classification methods in the biomedical datasets reported [144].	Supervised classification is the main goal, and the dataset contains multiple omics layers with matched samples	Strong predictive performance, but lower interpretability than factor-based methods and greater computational complexity
MOFA / MOFA+	Matrix factorization / factor analysis	Simulated datasets and TCGA multi-omics benchmarks	Have been found to be able to rediscover common structures in benchmarks, and the materials provided do not provide a MOFA particular numerical score [145].	Unsupervised exploration of heterogeneous bulk multi-omics datasets; identifying shared vs modality-specific variation	High interpretability and useful latent-factor structure, but may be less effective than nonlinear methods when interactions are highly complex
SNF	Network-based similarity fusion	Simulated CRC multi-omics and TCGA colorectal cancer datasets including DNA methylation, gene expression, and protein expression	Performed exceptionally as well as the most accurate in classifying the best in the CRC benchmark presented in the paper [146].	Patient stratification, subtype discovery, and similarity-based clustering across multiple modalities	Good for relationship-centered analysis, but less directly interpretable at the individual feature level
VAE	Deep learning / variational autoencoders	Single-cell parallel multi-omics datasets such as scRNA-seq + scATAC-seq and CITE-seq	Single cell multi omics benchmarks suggest that the performance depends on tasks and metrics, so deep VAE variants are one of the widely studied but no overall numeric advantage is demonstrated in the provided passages [28,147,148].	Large, nonlinear, multimodal datasets where denoising, latent representation learning, or	Flexible and powerful for nonlinear modeling, but usually less interpretable and more computationally demanding

				high-dimensional prediction is prioritized	
iCluster	Joint latent variable / matrix factorization	Simulations and TCGA breast cancer datasets	Able to recover shared structure with high accuracy in simulations, as well as generated competing clusterings on TCGA breast cancer data [145].	Integrative clustering and subtype discovery when interpretable latent structure is desired	More interpretable than deep learning approaches, but limited by stronger modeling assumptions
DIABLO (mixOmics)	Sparse PLS / supervised multi-omics integration	Multi-omics classification and biomarker discovery studies across matched cohorts	Supports correlated multi-omics feature selection and supervised discrimination across conditions [70].	Biomarker discovery and supervised classification when interpretable multi-omics signatures are needed	High interpretability and feature selection strength, but less suited for purely unsupervised structure discovery
WGCNA	Network / module-based integration	Transcriptomic datasets with epigenetic overlays such as accessibility or methylation-associated analyses	Useful for identifying co-regulated modules and relating them to epigenetic changes [85].	Linking chromatin accessibility or methylation shifts to coordinated gene-expression modules	Strong biological interpretability at the module level, but not designed for complex multimodal prediction
PANDA / NetBID	Regulatory-network inference	Multi-omics regulatory studies integrating transcriptomic and epigenetic signals	Useful for identifying candidate master regulators and hidden drivers from regulatory-network structure [86].	Regulatory network inference, transcription factor activity analysis, and mechanistic prioritization	High mechanistic value, but dependent on network assumptions and prior regulatory information
Seurat / Signac	Weighted nearest neighbor (WNN) analysis	Paired single-cell multi-omics such as 10x Multiome	Widely used for joint representation of RNA and chromatin accessibility in the same cells [57,90].	Paired single-cell RNA-seq and ATAC-seq integration, joint clustering, and peak-to-gene linkage	Highly practical and widely adopted, but mainly tailored to paired single-cell data

LIGER	Integrative non-negative matrix factorization (iNMF)	Unpaired or batch-affected single-cell multi-omics datasets	Effective for aligning datasets not measured in the same cells and for mitigating strong batch effects [91]	Unpaired single-cell integration or datasets generated across different platforms/labs	Useful for harmonization, though interpretation may be less direct than simpler factor-based bulk methods
OmiVAE / related autoencoder frameworks	Deep learning / autoencoder-based integration	Large cohort-level multi-omics datasets such as TCGA-style studies	Useful for latent compression and prediction in high-dimensional multi-omics settings [92]	Survival prediction, subtype prediction, or other nonlinear cohort-scale predictive tasks	Higher predictive flexibility, but reduced feature-level interpretability

 doi:10.20944/preprints202605.0428.v1

4.2. Matrix Factorization

The most established methods of multi-omics integration are matrix factorization (MF) approaches. These approaches are based on breaking down several data matrices into a collection of common and modality-specific latent factors [79,80]. MF models, including Multi-Omics Factor Analysis (MOFA) and Non-negative Matrix Factorization (NMF), are designed to identify low-dimensional representation of the highest variance of the various omics layers. These approaches can also be applied to find synchronized biological processes, including signaling pathway reflected in both DNA methylation and gene expression [79]. For single-cell data, Joint Semi-orthogonal Non-negative Matrix Factorization (JSNMF), enables the joint analysis of chromatin accessibility (scATAC-seq) and gene expression (scRNA-seq) [80]. JSNMF uses a consensus graph fusion strategy to deal with noise to modality and align the multiple omics layers more closely to allow accurate cell-type recognition [80]. Matrix factorization methods are highly interpretable, as the weights (or loadings) of the latent factors can often be directly associated with specific genes or epigenomic regions. These analyses provide clear biological insights into the underlying patterns captured by the model. They also tolerate lost data in any or all modalities.

However, MF methods normally depend on linear assumptions which cannot necessarily represent complex, nonlinear interactions between different layers of omics [81]. MOFA+ is especially effective for the unsupervised exploration of heterogeneous bulk multi-omics data, as its algorithm distinguishes between shared sources of variation and modality-specific factors. This capability allows researchers to more easily separate meaningful biological structures from technical noise. In contrast, DIABLO, a part of the mixOmics framework, is better suited for supervised contexts such as classification and biomarker discovery, where sparse cross-modal feature selection is essential. Ultimately, standard matrix factorization methods are most valuable when the primary goal is to discover latent structures, whereas sparse supervised variants are preferable when the priority is identifying interpretable multi-omics signatures associated with specific clinical phenotypes [69,70].

4.3. Network-Based Fusion

Network-based approaches assume that multi-omics data as a graph of interlinked nodes (samples or features) and leverage graph-theoretical methods to integrate and analyze the relationships across different omics layers. Similarity Network Fusion (SNF) constructs individual sample-similarity networks for each omics modality and recursively merges them into a single, comprehensive network. The method is especially useful for identifying coherent clusters of samples (e.g., disease subtypes) that are supported by multiple layers of molecular evidence [71].

Multi-omics Integration through Graph Convolutional Networks (MOGONET), is an enhanced network design. It performs Graph Convolutional Networks (GCNs) to learn omics-specific representations, treating samples as nodes within a patient similarity network. The final step of integration is performed using a View Correlation Discovery Network (VCDN), which identifies higher-level correlations in the label space between mRNA expression, DNA methylation and miRNA data [3]. Similarly, DeFusion is another integrative framework based on the denoised network regularization to combine multi-omics data. DeFusion enhances the strength of the biomarker discovery and classification tasks by removing technical noise and focusing on the shared topological structure of the networks [82,83].

Another important network-based method, TransNet (Transkingdom Network), is grounded in a systems biology framework, specially designed to identify causal components underlying host–microbiota interactions by integrating multi-omics data across different biological domains. The method combines various biological information from multiple kingdoms, including both host and microbial organisms to build extensive interrelationships within the biological system. The accurate analysis of these integrated networks opens the possibility of identifying some influential nodes or components that stimulate the host-microbiota dynamics. The approach is grounded in a comprehensive systems biology and computer network analysis of host-microbiota interactions,

 doi:10.20944/preprints202605.0428.v1

enabling the identification of key targets that can be leveraged for therapeutic interventions or mechanistic knowledge [84].

Other network-based approaches are especially helpful in situations where the objective is to deduce regulatory relationships rather than simply clustering samples. As an example, WGCNA (Weighted Gene Co-expression Network Analysis) can be applied to correlate transcriptional modules with epigenetic changes including chromatin accessibility or methylation changes. This enables researchers to connect global regulatory shifts directly to integrated gene-expression programs. Similarly, the PANDA (Platform for Architecture-Neutral Dynamic Analysis) and NetBid (Network-based Bayesian Inference of Drivers) frameworks are valuable for identifying candidate master regulators or the hidden drivers through the combination of regulatory-network structure with multi-omics measurements. Such approaches are particularly applicable when the analysis prioritizes mechanistic explanation and regulatory hierarchies over simple predictive classification [85,86].

4.4. Deep Learning Frameworks

Deep learning (DL) has completely transformed multi-omics analysis by offering adaptable architectures, capturing the most nonlinear interactions amongst high dimensional data. Variational Autoencoders (VAEs) and multimodal autoencoders are commonly applied in unsupervised integration. These models train compressed, nonlinear latent space that can simultaneously reconstruct multiple omics layers. This makes them exceptionally effective for data denoising and the imputation of missing values [72].

With the advent of spatial omics technologies, recent models, such as SpatialFuser, have been created to combine spatial transcriptomics with other modalities. SpatialFuser is a deep learning-based method designed to preserve spatial context during multi-omics data integration, enabling researchers to investigate how epigenetic regulation varies with tissue architecture [87]. Additionally, Multimodal Factorization AutoEncoder (MAE), combines multi-omics data with already known biological interaction networks. Using a combination of matrix factorization and deep autoencoders, MAE can find statistically significant, yet biologically meaningful features based on the available knowledge bases to enhance interpretability of the results [83,88].

Another sophisticated machine learning approach, ATHENA can find complex interactions between and among many levels of genomic data which correlate with cancer clinical outcomes, through Grammatical Evolution Neural Networks (GENN). This method mitigates the problems in predictions based on gene expression. These predictions often focus on single genes and overlook complex interactions or experiments involving different genomic layers like genome, epigenome, transcriptome and proteome. In a study involving ovarian cancer data from The Cancer Genome Atlas (TCGA), the ATHENA was demonstrated to identify both intra-level and inter-level genomic level interactions that can be predictive of patient survival. The application of ATHENA to integrate multiple levels of genomic data has been reported to achieve a balanced accuracy of 72.89% [89], outperforming models that relied on single-omics data. This suggests a synergistic effect, where the overall integration of multi-omics will be more appropriate in terms of depicting the multifaceted biological mechanisms that determine the outcomes of cancer. Ultimately, ATHENA aids in dissecting tumorigenesis and cancer progression by revealing complex molecular crosstalk, which facilitates the development of robust prognostic biomarkers and more personalized therapeutic strategies in oncology [89].

Further methodological attention is needed to single-cell multi-omics integration due to unique challenges of the data sparsity, high dimensionality, and the frequent lack of congruence between transcriptomic and chromatin accessibility profiles. When working with paired datasets, such as those generated by 10x Multiome, the Seurat/Signac framework provides a robust solution. Using weighted nearest neighbor (WNN) analysis, it enables joint clustering and the direct association of chromatin accessibility with gene expression at the single-cell level [90].

 doi:10.20944/preprints202605.0428.v1

In contrast LIGER is especially convenient with unpaired datasets or in systems with strong batch effects across different platforms. Its integrative non-negative matrix factorization (iNMF) algorithm is designed to align and harmonize datasets where measurements were not taken from the same individual cells [91]. Consequently, the selection of these tools depends not only on the omics modalities involved but also on whether the data is paired or unpaired and the degree of technical heterogeneity between experiments.

More generally, multimodal autoencoders and VAE-based models, like OmiVAE, can be most effectively applied when the task is to model nonlinear relationships of high complexity in large, high-dimensional cohort-level data sets. These methods are preferred when predictive performance—such as survival prediction or subtype classification—is the primary objective. However, they typically offer less direct biological interpretability compared to factor-based or network-based methods. Consequently, they are most appropriate when the dataset size is substantial and predictive accuracy is prioritized over feature-level explainability [72,92].

These integration families provide complementary strengths, as opposed to a mere hierarchy of performance. Statistical and matrix factorization methods have been more transparent and biologically interpretable and are thus appealing to hypothesis generation and mechanistic insight. In comparison, network-based and deep learning approaches can be more suitable in the context of nonlinear or high-order correlations among omics layers, though they can be computationally intensive and more difficult to interpret. The optimal approach is therefore determined by the specific research objective, whether it be exploratory analysis, subtype discovery, biomarker prioritization, or high-dimensional prediction [69,71,72,93].

These tools enable multi-omics data analysis, which utilizes statistical and machine learning algorithms, with potential to reveal previously hidden molecular interactions, and other biological insights that cannot be identified by analyzing the each omic layer separately [94,95]. MOFA (Multi-Omics Factor Analysis) model latent factors that drive variation in multiple datasets, can identify coordinated changes in the complexity of a biological sample [95]. Deep learning models can also improve the discovery of biomarkers, prognosis, and stratification of tumor subtypes by learning nonlinear relationships among epigenomic and multi-omics layers, such as transcriptomics, proteomics and clinical information [96]. Moreover, AI enables complete integration of epigenomic data with various omics modalities, which subsequently enhances focused oncology and tailored treatment plans. Integrative ML models, enable the development of sophisticated diagnostic and prognostic frameworks capable of predicting therapeutic response, reveal the mechanisms of resistance, and track the course of the disease over time [97].

4.5. Challenges and Practical Considerations in Multi-Omics Data Integration

Data harmonization is a key practical issue of all integration strategies. Multi-omics datasets also vary considerably in scale, distribution, sparsity, missingness, and platform-specific noise; therefore, they can only be directly compared after rigorous normalization and batch correction. These issues are particularly critical when integrating epigenomic assays, whereby assay-specific signal architecture and feature definitions of regulatory features often do not align naturally with transcriptomic or proteomic data. Consequently, integration methods must be selected not only for their predictive power but also for their ability to manage heterogeneous feature spaces, technical variability, and incomplete cross-modal data [12,71,98,99].

Despite recent developments, integration of epigenomics with other omics data has some outstanding challenges. One of the key challenges lies in the normalization of data across different platforms, given that different technologies produce widely varying data scales and noise characteristics, making direct integration and comparison difficult. The purpose of normalization is to ensure that the data in one dataset could be compared to the data in other datasets and minimize technical bias, although inconsistency usually persists [98,100]. Another important challenge is batch effects-systematic bias generated from variability in sample preparation, sequencing procedures, or

 doi:10.20944/preprints202605.0428.v1

laboratory conditions can hinder true biological variation. Addressing these batch effects without removing true biological signals, advanced computational methods are necessary [98,99].

Multi-omics data are also heterogeneous and high-dimensional, which poses significant challenges for interpretation. Frameworks (such as MOFA or iCluster) can identify integrated latent factors or clusters within multi omics data sets. These are often abstract constructs and require biological confirmation and practical knowledge. Their limited ability to be directly linked to underlying molecular mechanisms or disease processes further restricts the clinical applicability of these statistical patterns at the present stage [93,101].

Additionally, epigenomic data itself has modality-specific issues such as sparse signals, cellular heterogeneity, and intricate regulatory interactions, which further complicate their integration with other omics layers, including genomics, transcriptomics, and proteomics [12]. Therefore, integration approaches must be carefully designed to account for these distinctive features, enabling the extraction of biologically relevant findings.

5. Epigenomics in Disease and Cancer Biology: Clinical and Translational Applications

5.1. Epigenomics in Cancer: Mechanisms, Biomarkers, and Therapeutic Opportunities

Epigenomics is an important field of study in cancer and other diseases as it helps explain genome-wide changes like DNA methylation, histone modifications without altering the underlying DNA sequence. Epigenetic defects can be caused by genetic mutations that play a role in the process of tumorigenesis by reorganizing the gene expression, such as silencing tumor suppressor genes via hypermethylation of the CpG islands and abnormal acetylation of chromatin [102,103].

Recent research has demonstrated how the tumor microenvironment can reshape the epigenome, thereby contributing to cancer progression and metastasis [104]. Epigenetic alterations of DNA methylation, histone modification and RNA nucleotide modification (N6-methyladenosine, m6A) are involved in the progression of lung cancer, pulmonary diseases and can be studied using recent developments in epigenomic technologies [102,105]. Imbalances in small molecules associated with RNA epitranscriptomic machinery can disrupt RNA metabolism including transcription and stability, thereby contributing to uncontrolled cell proliferation and metastasis. Such advances in RNA sequencing technologies, allows high-resolution profiling of these RNA modifications, enabling researchers to gain deeper insights into their roles and potentially uncover novel therapeutic molecules in lung diseases [106].

One notable example is the application of an epigenomic analysis to cell-free tumor DNA in ovarian cancer, where epigenomic profiles of DNA methylation combined with artificial intelligence (AI) achieved near-perfect diagnostic value (AUC up to 1.00), exceeding accuracy of conventional serum markers and imaging [107]. The identified epigenomic signatures offer biologically realistic interpretations of cancer-related pathways, providing diagnostics reliability than conventional morphological or protein-based approaches [107]. Additionally, Kim and Kim, 2009 reported that epigenetic markers in urine of patients with bladder cancer exhibit higher sensitivity than conventional cytology for bladder cancer detection. This enhanced sensitivity arises from the ability to detect promoter hypermethylation of tumor suppressor genes, which are linked with tumor development and prognosis [108].

Epigenomics also offers new prospects in disease prevention and personalized medicine. Isothiocyanates and natural substances with epigenetic-modifying properties, have been examined in their capacity to prevent the early-stage of cancers including skin, colon, lung, breast, and prostate cancers. These agents provide antioxidative, anti-inflammatory, and epigenetic effects, potentially slowing cancer progression and metastasis by modulating epigenomic machinery [109].

Recent research suggests that noncanonical histone modifications also play a role in the renal protection. Histone β-hydroxybutyrylation has been demonstrated to act as the mechanism through which β-hydroxybutyrate produces reno-protective effects in the Dahl rat, which demonstrates a mechanistic association between metabolic state and epigenetic regulation in kidney disease [110].

 doi:10.20944/preprints202605.0428.v1

Similarly, epigenetic drugs that target DNA methylation and histone modification are under development to be applied in therapeutic management of cancers like cervical cancer. These drugs are aimed to enable earlier diagnosis, provide rich and functional molecular information on cancer states, decrease dependency on invasive sampling and subjective interpretation, and provide better patient stratification and disease surveillance [1].

BRD9 (Bromodomain-containing protein 9), as an epigenetic reader, can be targeted by drugs. Suppressing BRD9 minimizes pigmentation, gene expression and melanin synthesis. This suggests that targeting bromodomain proteins may be useful for therapeutics [111]. These characteristics will be a major improvement upon conventional diagnostic modalities and are potentially transforming precise oncology [107,108,112].

Techniques to study single-cell epigenomics have further enhanced the overall understanding of cancer biology, allowing insights into cancer stage progression, treatment response and mechanisms of relapse by elucidating epigenetic heterogeneity within tumors. These techniques capture chromatin phenotypes such as accessibility, histone reactions, as well as DNA methylation at the single-cell scale, which give a more comprehensive and dynamic perspective on tumor development [113].

Epigenomic signatures, including DNA methylation patterns, have also been effectively applied to identify the tissue of origin in the cancers of unknown primary (CUP), a clinically significant phenotype, as they improve diagnostic specificity and guide therapeutic decision-making [114]. The disrupted genes and pathways through epigenetic changes can elucidate the tumorigenesis processes and have been used to develop specific epigenetic therapeutics, supporting patient stratification and disease monitoring, including the detection of minimal residual disease [115,116]. Epigenomic knowledge is highly valuable for precision medicine, as integrating epigenomic and genomic data enables the tailoring of therapies to specific molecular patterns, thereby enhancing therapeutic effectiveness and clinical impact [117]. Minimally invasive approaches for cancer screening, early-detection, tissue-of-origin, and disease monitoring technologies have become possible due to the development of cell-free DNA (cfDNA) technologies, mainly based on epigenetic signatures (primarily DNA methylation) in fluid bio-samples [116]. Another benefit of epigenomic profiling is that it avoids the shortcomings of circulating tumor DNA approaches, which primarily capture genetic mutations but often fail to reflect dynamic transcriptional programs and epigenetic modifications that regulate cancer phenotypes. Immunoprecipitation-based methods capable of detecting histone modifications and DNA methylation in liquid biopsies enable the identification of drug targets, resistance mechanisms and cancer subtype classification using only plasma samples. This gives important information not accessible through conventional tissue biopsies or genetic testing approaches [112].

As an example, 5hmC profiling of cfDNA has become a novel instrument of oncology with precision because it mirrors cancer biology, such as initiation, progression, and metastasis in diverse types of cancer. Genome-wide 5hmC profiling provides sensitive screening, diagnostics, and prognostic analysis of cancer, wherever traditional methods are only limited to tissue biopsies [118,119]. The identification of epigenetic mutations along with approval of epigenetic drugs like DNA methyltransferase (DNMT) and histone deacetylase (HDAC) inhibitors, represents a paradigm shift in the therapeutics field. These epigenetic drugs target aberrant epigenetic mechanisms and are undergoing extensive clinical evaluation across multiple cancer types [1,120].

Moreover, epigenomics has provided insights into how environmental exposures can contribute to cancer development through the deregulation of epigenetic mechanisms. These insights have driven the introduction of new epigenetic cancer therapeutics that target enzymes responsible for writing, erasing, or reading DNA and histone modifications, including DNA methyltransferases, histone acetyltransferases, deacetylases, and demethylases. An illustration of the therapeutic potential of reversible epigenetic modifications in cancer is the growing number of epigenetic drugs that have already been approved in a clinical setting [1,121,122].

 doi:10.20944/preprints202605.0428.v1

5.2. Epigenomics Beyond Cancer: Applications in Other Diseases

Beyond cancer, epigenomics plays a crucial role in other diseases, including stroke and cardiometabolic disorders, by influencing gene regulation and contributing to disease onset, progression, and patient outcomes. Epigenetic changes, such as DNA methylation and histone acetylation, are associated with the etiology of stroke. Understanding these changes offers valuable opportunities to develop novel preventive and therapeutic interventions [123]. Environmental and lifestyle alterations can also cause epigenetic alterations in cardiometabolic disease, which influence the presentation of the disease. The deciphering of these epigenetic signatures may reveal novel molecular targets for precision medicine, enabling strategies to minimize cardiovascular risk [124].

One of the most significant advances has been the development of epigenetic clocks based on DNA methylation patterns. These tools have enhanced the molecular-level investigation of aging processes and age-related diseases, providing more precise biomarkers of biological age. Moreover, these tools provide insight into how environmental exposures such as nutrient imbalances, particularly in B-vitamins, and polymorphisms in one-carbon metabolism pathways impact epigenetic aging and disease susceptibility, highlighting the intersection of nutritional epigenomics and health outcomes [125].

Advances in chromatin biology in neuroscience have led to changes in epigenetic processes. These processes can regulate the expression of genes in neurons, and this is important for synaptic plasticity and memories. In addition to the histone phosphorylation and acetylation, recent findings have highlighted that DNA and histone methylation also play crucial roles in behavior and long-term memory regulation [126]. The study of the interactions between epigenetic remodeling and transcription factor activity can provide important clues into the mechanisms by which transcriptional regulation governs memory storage [126]. In the cardiovascular biology of heart failure, recent research indicates that chromatin-modifying enzymes and 3D structure of the genome, including the lamina-associated domains, play important roles in the epigenetic regulation of gene expression, influencing disease pathology. Recent progress has revealed that histone demethylases and deacetylation play key roles in regulating fetal gene expression programs and cardio protection and that epigenetic responses to stress and external stimuli can determine whether the outcome is adaptive or pathological [127].

Beyond cancer, the translational research in epigenomics has also been applied to mental health and developmental diseases, where epigenetic processes play critical roles in brain development and function [128].

The combination of SNPs and transcriptomics along with DNA methylation has been especially beneficial for metabolic studies. In a recent study, the DIABLO (Data Integration Analysis for Biomarker discovery using Latent cOmponents) algorithm (of the mixOmics R package) was employed to accurately predict Type 2 Diabetes (T2D) using several omics layers from human pancreatic islets. This integrative examination identified novel candidate biomarkers, including the SACS (Sacsin Molecular Chaperone), TXNIP (Thioredoxin-interacting protein) DNA methylation and OPRD1(Opioid Receptor Delta 1) expression, that offer understanding of biological processes of mitochondrial malfunctioning and altered insulin secretion in T2D [129].

Epigenetic clocks - models that estimate biological age or health condition based on DNA methylation pattern have achieved next levels of resolution and accuracy. BS-clock is a recently developed DNA methylation clock optimized using high-resolution bisulfite sequencing information, enabling more precise predictions of human aging compared to the earlier array-based models [130]. Moreover, iTARGET (Interpretable Tailored Age Regression of Grouped Epigenetic Traits) employs explainable boosting machines to generate interpretable age predictions, revealing age-specific CpG interactions that were previously hidden [131]. These predictors are increasingly applicable in personalized medicine and can be used as biomarkers of disease risk and environment exposure [132]. Moreover, recent research shows that bioinformatic tools can help identify therapeutic targets for certain chronic diseases. As an example, DNA methylation of SACS in Type 2 Diabetes has been suggested as a potential target for restoring mitochondrial function in future [129].

 doi:10.20944/preprints202605.0428.v1

Regarding respiratory diseases, DNA methylation changes associated with asthma have been identified using tools like easyEWAS, which facilitate the epigenetic analysis of how environmental exposures impact lung health [31] .

5.3. Clinical Translation: Current Progress, Standardization, and Implementation Challenges

There are clinical and translational applications of epigenomics important in advancing the process of diagnosis, prognosis, and personalized therapy in various medical fields. One of the most significant challenges in the rapidly evolving epigenetics field is standardization, which has been addressed through several updated or newly developed R and Bioconductor packages. For example, easyEWAS is a convenient software, which combines the latest approaches to Epigenome-Wide Association Studies, that can account either the position-specific or region-specific analysis of differential methylation [53]. The BISulfite-seq Command line User Interface Toolkit (with the biscuiteer R package) is a standards-compliant genetic and epigenetic dual inference suite available for use in bulk and single-cell analyses [133]. Moreover, TENET is an epigenetic-based network tracing algorithm that determines key regulatory entities and transcription factors across ten cancer types in pan-cancer studies [134].

Overall, epigenomics has increasingly been transforming clinical practice, particularly in precision oncology, and allowing better diagnostics, personalized therapies, non-invasive monitoring, and biomarker discovery. Further development and connection with multi-omics data ensures that the sphere of its implementation will continue to grow in various fields, pushing the development of translational medicine.

Despite these promising uses in early cancer detection, personalized medicine, and biological age estimation, the field of epigenomics faces significant hurdles that must be addressed to ensure successful clinical translation. Reports of exceptionally high diagnostic or prognostic performance derived from small patient cohorts should be interpreted with caution unless validated in large, diverse, and independent datasets. In practice, clinical adoption depends on multiple factors, including robust model performance, assay standardization, and reproducibility across laboratories. Additional considerations include sample quality, affordability, turnaround time, scalability for routine use, effective control of batch effects, and overall cost-effectiveness. Equally important is the compatibility of these approaches with existing clinical workflows. Moreover, regulatory approval, prospective validation in real-world clinical settings, and the limited interpretability of certain machine-learning models remain significant barriers to implementation in clinical decision-making. Therefore, the path from computational epigenomic discovery to clinical application requires not only the identification of reliable biomarkers but also rigorous technical validation, multicenter studies, and clear demonstration of clinical utility beyond retrospective performance metrics [135–137].

7. Discussion- Current Limitations and Future Directions

The analysis of epigenomic data is associated with various problems due to the complexity and variability of epigenetic processes, technical and computational constraints. The integration and interpretation of large volumes of data produced by high-throughput sequencing technologies that can map DNA methylation, histone states, chromatin accessibility, and other epigenetic states on a genome-wide scale is one of the key challenges. These data tend to be high-dimensional, sparse, and heterogeneous and need advanced computing resources to derive any meaningful biological information [12,149]. Another challenge is normalization of epigenomic data, because of the differences between sequencing depth and signal-to-noise ratios among replicate experiments or cell types. Current normalization procedures might not be able to explain these variations simultaneously, which could cause the obscuration of true positive biological variation (Figure 5). More modern approaches have been designed to address these problems, such as S3norm, which normalize both sequencing depth and signal differences to enhance comparability of epigenomic datasets [13].

 doi:10.20944/preprints202605.0428.v1

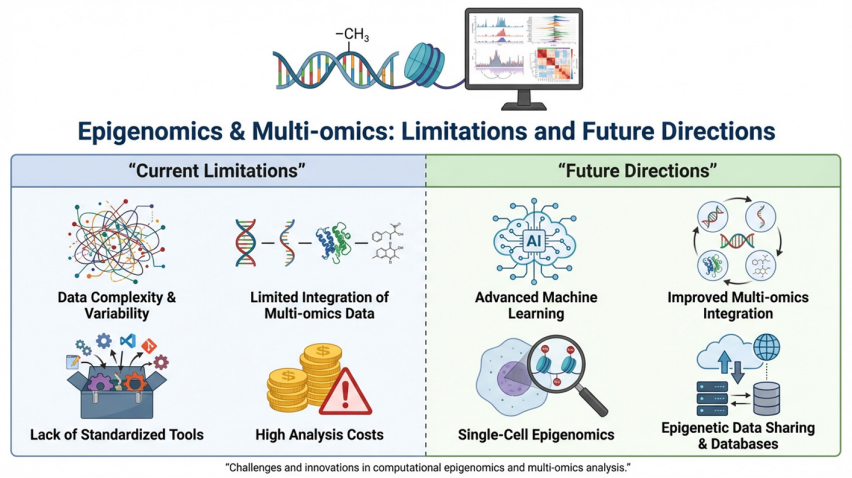


Figure 5. Current Challenges and Future Horizons in Epigenomics. A comparative overview highlighting the existing bottlenecks in epigenomic research versus the emerging strategies designed to overcome them. The figure summarizes the field's evolution from high-cost, fragmented analysis toward a more integrated and accessible future.

Epigenomic regulation is also biologically complex, posing major interpretative challenges. Numerous epigenetic changes occur outside well-annotated protein-coding regions, which makes the functional annotation of regulatory elements difficult. Combination of epigenomic data sets with 3D genome structure, enrichment analysis and machine learning approaches are actively pursued to implicate functional significance; however, a detailed and assured reading is intricate [149]. Causality or association in epigenetic studies is also difficult to measure, particularly in the light of interaction between genetic and epigenetic variables influencing variation at the population level. It can be seen in the research of diseases and traits, where it is very difficult to separate the influence of epigenetics and genetic variation that forms the background [150]. Epigenomic research is also complicated by technological and ethical issues. Quality of samples used in an experiment and assay biases are experimental factors that influence the reliability of data. Furthermore, the problem of ethical concerns regarding privacy, possible discrimination due to epigenetic data, and fair access to clinical uses of epigenetics are important questions regarding the translation of epigenomics into personalized medicine [151].

The analysis of epigenomic data in the disease setting, including cancer, is complicated by tumor heterogeneity and is required to develop computational approaches that can discover biomarkers and classify patients with complex DNA methylation patterns. Deep learning methods are promising for epigenomic analysis, but they are associated with interpretability and issues regarding data demands [152].

Artificial intelligence (AI) has revolutionized epigenomic data analysis by providing computational models capable of simulating the scale, complexity, and regulatory interdependence inherent in epigenetic systems. Critical literature suggests that traditional analyses—often reliant on linear models and manual feature engineering—fail to capture the intricate relationships between DNA methylation, histone modifications, chromatin accessibility, and 3D genome organization [153,154]. Conversely, modern machine learning (ML) and deep learning (DL) architectures, such as CNNs, GNNs, and transformer-based models, can detect patterns directly from raw sequence and epigenomic data. These frameworks facilitate the precise estimation of chromatin states, promoter-enhancer interactions, and transcriptional activity [96,154]. Furthermore, the integration of unsupervised and self-supervised learning has enabled the discovery of novel regulatory programs without the need for extensive labeled datasets, maximizing the utility of large-scale consortia data [153].

 doi:10.20944/preprints202605.0428.v1

Despite these advancements, several major challenges continue to hinder the clinical integration of AI-based epigenomic tools. Model interpretability remains a primary obstacle; many deep learning frameworks function as 'black boxes' that fail to provide the biological insights necessary to instill clinician confidence in diagnostic or prognostic results [153,155]. Furthermore, a lack of generalizability and inherent data bias often arise because models are trained on datasets unrepresentative of diverse populations, cancer subtypes, or disease stages. This frequently leads to degraded performance when tools are applied to independent cohorts or real-world clinical contexts [153].

Recent research aims to mitigate these shortcomings through explainable AI (XAI) tools—such as feature attribution and attention mechanisms—and biologically grounded architectures that enhance transparency and ensure model predictions are biologically relevant [155]. Standardized benchmarking protocols and the use of large-scale consortia, such as ENCODE, TCGA, and Roadmap Epigenomics, are essential for improving reproducibility and enabling fair comparisons across studies [153]. Finally, rigorous clinical validation, multi-center trials, and external cohort assessments are required to demonstrate the robustness and utility necessary for regulatory approval [135].

8. Methodology of Literature Search

This study was conducted as a narrative review to provide a critical and representative overview of current bioinformatics approaches to epigenomic and multi-omics research. In this study, we conducted a comprehensive literature review spanning the last decade using electronic databases, including PubMed and Google Scholar. Additionally, we examined existing review articles to discover further research in single-cell epigenomics and the integration of epigenomics with other omics modalities. Articles were filtered out if they were not published in English, not directly related to epigenomic analysis or computational integration or lacked sufficient methodological detail for meaningful discussion.

The following keywords and their combinations were used: "Epigenomics", "Histone modification", "DNA methylation", "ChIP-seq", "ATAC-seq", "Bioinformatics tools", "Epigenetic regulation", "Multi-omics integration", "Single-cell epigenomics", "Machine learning", "epigenetic biomarkers", "scATAC-seq", "Principal component analysis (PCA) and canonical correlation analysis (CCA)", "Matrix Factorization", "Computational integration strategies". The selection of studies was restricted by relevance to the area of interest of this review. We prioritized peer-reviewed articles detailing popular experimental platforms, developed or developing computational tools, and integration systems, and clinical uses in health and disease. Where needed, landmark studies were included to provide historical context, but more recent studies were given priority to represent the current developments in the field.

Since this is a narrative review, a literature selection was not intended to be exhaustive, and no PRISMA-style screening framework was utilized. This method can lead to selection bias especially on highly cited, English-language, and more readily available studies. Moreover, the use of PubMed and Google Scholar might not be able to retrieve all the relevant publications that are indexed in other databases. However, the search strategy was aimed at offering a balanced and recent overview of key approaches, tools, issues, and future perspectives of integrative epigenomics.

9. Conclusions

Advances in epigenomics are primarily driven by the evolution of experimental technologies and increasingly sophisticated computational methods for data processing and interpretation. A central conclusion of this review is that no single methodology is universally optimal. Instead, the selection and integration of tools must be tailored to the specific assay type, data structure, and biological objective, while maintaining a requisite balance between interpretability, scalability, and predictive performance. Nevertheless, persistent challenges—including issues with reproducibility,

 doi:10.20944/preprints202605.0428.v1

normalization, non-coding region annotation, platform harmonization, and clinical validation—remain significant barriers to robust biological discovery.

The translation of epigenomic knowledge into clinical practice, particularly within precision oncology and biomarker development, demonstrates its transformative impact on modern medicine. Emerging frontiers, such as single-cell epigenomics, multi-omics integration, and explainable AI (XAI), are shifting the field from descriptive profiling toward mechanistic and clinically actionable models of gene regulation. This review underscores that future progress requires more than the mere accumulation of data; it necessitates the development of transparent, benchmarked, and biologically informed computational frameworks. Only through such rigor can these systems facilitate reproducible discoveries that are reliable enough for routine clinical implementation.

Author Contributions: Conceptualization, SB; data curation, SB, SA; writing—original draft, SB, LB; writing—review and editing, SB, SA, LB, AEB; visualization, SB, SA, LB, AEB; supervision, LB, and AEB. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: We thank the section managing editor for recommendation of the article to the featured paper in Epigenome Journal. Because of space restrictions, we cannot cite many important pieces of recent literature in this field. We apologize to all those colleagues whose work contributed significantly but could not be included here. During the preparation of this review, the author SA used drawio and FigureLabs for the purposes of designing the figures. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ATAC-seq - Assay for Transposase-Accessible Chromatin using sequencing
ATHENA - Analysis Tool for Heritable and Environmental Network Associations
BRD9 - Bromodomain-containing protein 9
CUP - Cancers of Unknown Primary
CfDNA - Cell-free DNA
CoBRA - Combined Bisulfite Restriction Analysis
ChIP-seq - Chromatin Immunoprecipitation sequencing
CCA - Canonical Correlation Analysis
casTLE - Cas9 High-Throughput Maximum Likelihood Estimator
DROMPA - Discrepancy Restricted Overlapping Mapped Peak Analyzer
DL - Deep Learning
DNMT - DNA methyltransferase
DIABLO - Data Integration Analysis for Biomarker discovery using Latent cOmponents
FAIRE-Seq - Formaldehyde-Assisted Isolation of Regulatory Elements
GENN - Grammatical Evolution Neural Networks
GREAT - Genomic Regions Enrichment of Annotations Tool
GCNs - Graph Convolutional Networks
HDAC - Histone Deacetylase
HOMER - Hypergeometric Optimization of Motif EnRichment

 doi:10.20944/preprints202605.0428.v1

HMMRATAC - Hidden Markov ModeleR for ATAC
iTARGET - Interpretable Tailored Age Regression of Grouped Epigenetic Traits
JSNMF - Joint Semi-orthogonal Non-negative Matrix Factorization
LIANTI - Linear Amplification via Transposon Insertion
MeDIP-seq – Methylated DNA Immunoprecipitation Sequencing
MSP - Methylation-Specific Polymerase Chain Reaction
MAPS - Model-based Analysis of PLAC seq
MALBAC - Multiple Annealing and Looping-Based Amplification Cycles
MF - Matrix Factorization
MOFA - Multi-Omics Factor Analysis
MAE - Multimodal Factorization AutoEncoder
MOGONET - Multi-omics Integration through Graph Convolutional Networks
NMF - Non-negative Matrix Factorization
OPRD1- Opioid Receptor Delta 1
PCA - Principal Component Analysis
RRBS - Reduced Representation Bisulfite Sequencing
scWGBS - Single-Cell Whole Genome Bisulfite Sequencing
scRRBS - Single-Cell Reduced Representation Bisulfite Sequencing
scBS-seq - Single-Cell Bisulfite Sequencing
snmC-seq - Single-Nucleus Methylcytosine Sequencing
scATAC-seq - Single-Cell ATAC-seq
ScDEC-Hi-C - Single-Cell deep embedded clustering
SCAVENGE - Single-Cell Analysis of Variant Enrichment through Network propagation of GEnomic data
scCASE - Single-Cell Chromatin Accessibility Sequencing Enhancement
SDGCCA - Supervised Deep Generalized Canonical Correlation Analysis
SNF - Similarity Network Fusion
SACS - Sacsin Molecular Chaperone
SEACR — Sparse Enrichment Analysis for CUT&RUN
TCGA - The Cancer Genome Atlas
TransNet - Transkingdom Network
TXNIP - Thioredoxin-interacting protein
VCDN - View Correlation Discovery Network
VAEs - Variational Autoencoders
WGBS - Whole Genome Bisulfite Sequencing

References

1. Kabekkodu SP, Chakrabarty S, Ghosh S, Brand A, Satyamoorthy K. Epigenomics, pharmacoepigenomics, and personalized medicine in cervical cancer. Public Health Genomics. 2017;20(2):100-15.
2. Novik K, Nimmrich I, Genc B, Maier S, Piepenbrock C, Olek A, et al. Epigenomics: genome-wide study of methylation phenomena. Current issues in molecular biology. 2002;4(4):111-28.
3. Scherer M, Schmidt F, Lazareva O, Walter J, Baumbach J, Schulz MH, et al. Machine learning for deciphering cell heterogeneity and gene regulation. Nature Computational Science. 2021;1(3):183-91.
4. Schwartzman O, Tanay A. Single-cell epigenomics: techniques and emerging applications. Nature Reviews Genetics. 2015;16(12):716-26.
5. Arnau-Soler A, Tremblay BL, Sun Y, Madore AM, Simard M, Kersten ET, et al. Food Allergy Genetics and Epigenetics: A Review of Genome-Wide Association Studies. Allergy. 2025;80(1):106-31.
6. Baccarelli AA, Ordovás J. Epigenetics of early cardiometabolic disease: mechanisms and precision medicine. Circulation research. 2023;132(12):1648-62.
7. Gemmati D, Villanova M, Scarpellini F, Milani D, Cecchi R, Singh AV, et al. Epigenetic modifications and transgenerational inheritance in women victims of violence (EWVV). Environmental Epigenetics. 2025;11(1):dvaf025.

 doi:10.20944/preprints202605.0428.v1

8. Vaidya H, Jelinek J, Issa J-PJ. DNA methylation, aging, and cancer. Epigenomes. 2025;9(2):18.
9. Fingerman IM, Hassan, T., Jiang, Z., Ratzat, W., Mcdaniel, L., Schuler, G. D., Zhang, X., & Cohen, R. F. NCBI Epigenomics: a new public resource for exploring epigenomic data sets. . Nucleic Acids Research, 39(Database issue), D908–D912 https://doiorg/101093/nar/gkq1146. 2010.
10. Han Y, He X. Integrating epigenomics into the understanding of biomedical insight. Bioinformatics and Biology insights. 2016;10:BBI. S38427.
11. Albrecht F, List M, Bock C, Lengauer T. DeepBlue epigenomic data server: programmatic data retrieval and analysis of epigenome region sets. Nucleic acids research. 2016;44(W1):W581-W6.
12. Danese A, Richter ML, Chaichoompu K, Fischer DS, Theis FJ, Colomé-Tatché M. EpiScanpy: integrated single-cell epigenomic analysis. Nature communications. 2021;12(1):5228.
13. Xiang G, Keller CA, Giardine B, An L, Li Q, Zhang Y, et al. S3norm: simultaneous normalization of sequencing depth and signal-to-noise ratio in epigenomic data. Nucleic acids research. 2020;48(8):e43-e.
14. Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. Imeta. 2023;2(2):e107.
15. Chen S, Huang T, Zhou Y, Han Y, Xu M, Gu J. AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. BMC bioinformatics. 2017;18(Suppl 3):80.
16. Zhang Y, Hardison RC. Accurate and reproducible functional maps in 127 human cell types via 2D genome segmentation. Nucleic acids research. 2017;45(17):9823-36.
17. He X, Chen S, Li R, Han X, He Z, Yuan D, et al. Comprehensive fundamental somatic variant calling and quality management strategies for human cancer genomes. Briefings in Bioinformatics. 2021;22(3):bbaa083.
18. Suryatenggara J, Yong KJ, Tenen DE, Tenen DG, Bassal MA. ChIP-AP: an integrated analysis pipeline for unbiased ChIP-seq analysis. Briefings in bioinformatics. 2022;23(1):bbab537.
19. Guo X, Chen Q, Zhang Y, Zhang Y, Liu Y, Duan S, et al. Methylation reference datasets from quartet DNA materials for benchmarking epigenome sequencing. Nature Communications. 2025;16(1):9202.
20. Jeon H, Lee H, Kang B, Jang I, Roh T-Y. Comparative analysis of commonly used peak calling programs for ChIP-Seq analysis. Genomics & informatics. 2020;18(4):e42.
21. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: a method for assaying chromatin accessibility genome-wide. Current protocols in molecular biology. 2015;109(1):21.9. 1-.9. 9.
22. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Kheradpour P, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317.
23. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. nature. 2009;462(7271):315-22.
24. Park PJ. ChIP–seq: advantages and challenges of a maturing technology. Nature reviews genetics. 2009;10(10):669-80.
25. Benčik I, Saftić Martinović L, Mladenić T, Ostojić S, Dević Pavlić S. From DNA Methylation and Histone Modifications to Non-Coding RNAs: Evaluating Tools for Epigenetic Research. Applied Sciences. 2025;15(18):9940.
26. Neefs I, Ibrahim J, Peeters M, Van Camp G, Op de Beeck K. The technology landscape for detection of DNA methylation in cancer liquid biopsies. Epigenetics. 2025;20(1):2453273.
27. Yang Y, Wen X, Wang L. Advancements in DNA methylation technologies and their application in cancer diagnosis. Epigenetics. 2025;20(1):2539995.
28. Yang X, Mann KK, Wu H, Ding J. scCross: a deep generative model for unifying single-cell multi-omics with seamless integration, cross-modal generation, and in silico exploration. Genome Biology. 2024;25(1):198.
29. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. bioinformatics. 2011;27(11):1571-2.
30. Cheng S, Miao B, Li T, Zhao G, Zhang B. Review and evaluate the bioinformatics analysis strategies of ATAC-seq and CUT&tag data. Genomics, Proteomics & Bioinformatics. 2024;22(3):qzae054.
31. Wang R, Wu Y, Zhou Z, Ma Y, Zhang W, Wang Z, et al. Benchmark of chromatin–protein interaction methods in haploid round spermatids. Frontiers in Cell and Developmental Biology. 2025;13:1572405.

 doi:10.20944/preprints202605.0428.v1

32. Xiong C, Wang J, Li X, Li G, Wang X. The development and application of cleavage under targets and tagmentation (CUT&Tag) technology. Journal of Biological Methods. 2024;11(3):e99010019.
33. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, et al. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nature biotechnology. 2009;27(1):66.
34. Wilbanks EG, Facciotti MT. Evaluation of algorithm performance in ChIP-seq peak detection. PloS one. 2010;5(7):e11471.
35. Nakato R, Itoh T, Shirahige K. DROMPA: easy-to-handle peak calling and visualization software for the computational analysis and validation of ChIP-seq data. Genes to Cells. 2013;18(7):589-601.
36. Qiu X, Feit AS, Feiglin A, Xie Y, Kesten N, Taing L, et al. CoBRA: containerized bioinformatics workflow for reproducible ChIP/ATAC-seq analysis. Genomics, proteomics & bioinformatics. 2021;19(4):652-61.
37. Lerdrup M, Johansen JV, Agrawal-Singh S, Hansen K. An interactive environment for agile analysis and visualization of ChIP-sequencing data. Nature structural & molecular biology. 2016;23(4):349-57.
38. Eder T, Grebien F. Comprehensive assessment of differential ChIP-seq tools guides optimal algorithm selection. Genome Biology. 2022;23(1):119.
39. Juric I, Yu M, Abnousi A, Raviram R, Fang R, Zhao Y, et al. MAPS: Model-based analysis of long-range chromatin interactions from PLAC-seq and HiChIP experiments. PLoS computational biology. 2019;15(4):e1006982.
40. Meers MP, Tenenbaum D, Henikoff S. Peak calling by Sparse Enrichment Analysis for CUT&RUN chromatin profiling. Epigenetics & chromatin. 2019;12(1):42.
41. Yashar WM, Kong G, VanCampen J, Curtiss BM, Coleman DJ, Carbone L, et al. GoPeaks: histone modification peak calling for CUT&Tag. Genome biology. 2022;23(1):144.
42. Nooranikhojasteh A, Tavallaee G, Orouji E. Benchmarking peak calling methods for CUT&RUN. Bioinformatics. 2025;41(7):btaf375.
43. Shpynov O, Artyomov MN. Accurate chromatin marks peak calling with Omnipeak. Nucleic acids research. 2026;54(1):gkaf1454.
44. Lin L, Liu Y. Advances in epigenomic sequencing and their applications in cancer diagnostics. Precision Chemistry. 2025;3(10):581-603.
45. Ahmed Z, Ucar D. I-ATAC: interactive pipeline for the management and pre-processing of ATAC-seq samples. PeerJ. 2017;5:e4040.
46. Smith JP, Sheffield NC. Analytical approaches for ATAC-seq data analysis. Current protocols in human genetics. 2020;106(1):e101.
47. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. Science. 2018;362(6413):eaav1898.
48. Lu RJ-H, Liu Y-T, Huang CW, Yen M-R, Lin C-Y, Chen P-Y. ATACgraph: profiling genome-wide chromatin accessibility from ATAC-seq. Frontiers in Genetics. 2021;11:618478.
49. Tarbell ED, Liu T. HMMRATAC: a Hidden Markov ModeleR for ATAC-seq. Nucleic acids research. 2019;47(16):e91-e.
50. Yan F, Powell DR, Curtis DJ, Wong NC. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. Genome biology. 2020;21(1):22.
51. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841-2.
52. Wang Q, Li M, Wu T, Zhan L, Li L, Chen M, et al. Exploring epigenomic datasets by ChIPseeker. Current protocols. 2022;2(10):e585.
53. Yu G, Wang L-G, He Q-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics. 2015;31(14):2382-3.
54. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nature biotechnology. 2010;28(5):495-501.
55. Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, Chang HY, et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nature genetics. 2021;53(3):403-11.

 doi:10.20944/preprints202605.0428.v1

56. Kwok AWC, Shim H, McCarthy DJ. A hierarchical, count-based model highlights challenges in scATAC-seq data analysis and points to opportunities to extract finer-resolution information. Genome Biology. 2025;26(1):282.
57. Stuart T, Srivastava A, Madad S, Lareau CA, Satija R. Single-cell chromatin state analysis with Signac. Nature methods. 2021;18(11):1333-41.
58. Farlik M, Sheffield NC, Nuzzo A, Datlinger P, Schönegger A, Klughammer J, et al. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. Cell reports. 2015;10(8):1386-97.
59. Guan A, Quek C. Single-cell multi-omics: insights into therapeutic innovations to advance treatment in cancer. International Journal of Molecular Sciences. 2025;26(6):2447.
60. Lim J, Park C, Kim M, Kim H, Kim J, Lee D-S. Advances in single-cell omics and multiomics for high-resolution molecular profiling. Experimental & molecular medicine. 2024;56(3):515-26.
61. Plummer D, Lang X, Zhang S, Li Y, Li J, Jin F. A comprehensive benchmark of single-cell Hi-C embedding tools. Nature Communications. 2025;16(1):9119.
62. Kopp W, Akalin A, Ohler U. Simultaneous dimensionality reduction and integration for single-cell ATAC-seq data using deep learning. Nature Machine Intelligence. 2022;4(2):162-8.
63. Chen H, Lareau C, Andreani T, Vinyard ME, Garcia SP, Clement K, et al. Assessment of computational methods for the analysis of single-cell ATAC-seq data. Genome biology. 2019;20(1):241.
64. Tang S, Cui X, Wang R, Li S, Li S, Huang X, et al. scCASE: accurate and interpretable enhancement for single-cell chromatin accessibility sequencing data. Nature communications. 2024;15(1):1629.
65. Lance C, Luecken MD, Burkhardt DB, Cannoodt R, Rautenstrauch P, Laddach A, et al. Multimodal single cell data integration challenge: results and lessons learned. BioRxiv. 2022:2022.04. 11.487796.
66. Li S, Hua H, Chen S. Graph neural networks for single-cell omics data: a review of approaches and applications. Briefings in bioinformatics. 2025;26(2):bbaf109.
67. Cui X, Chen X, Li Z, Gao Z, Chen S, Jiang R. Discrete latent embedding of single-cell chromatin accessibility sequencing data for uncovering cell heterogeneity. Nature computational science. 2024;4(5):346-59.
68. Sibilio P, De Smaele E, Paci P, Conte F. Integrating multi-omics data: Methods and applications in human complex diseases. Biotechnology Reports. 2025:e00938.
69. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome biology. 2020;21(1):111.
70. Rohart F, Gautier B, Singh A, Lê Cao K-A. mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS computational biology. 2017;13(11):e1005752.
71. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: mathematical aspects. BMC bioinformatics. 2016;17(Suppl 2):S15.
72. Tsimenidis S, Vrochidou E, Papakostas GA. Omics data and data representations for deep learning-based predictive modeling. International Journal of Molecular Sciences. 2022;23(20):12272.
73. Han X. Nonnegative principal component analysis for cancer molecular pattern discovery. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2009;7(3):537-49.
74. Jiang M-Z, Aguet F, Ardlie K, Chen J, Cornell E, Cruz D, et al. Canonical correlation analysis for multi-omics: Application to cross-cohort analysis. PLoS genetics. 2023;19(5):e1010517.
75. Moon S, Hwang J, Lee H. SDGCCA: supervised deep generalized canonical correlation analysis for multi-omics integration. Journal of Computational Biology. 2022;29(8):892-907.
76. Xu M, Zhu Z, Zhang X, Zhao Y, Li X. Canonical correlation analysis with l 2, 1-norm for multiview data representation. IEEE transactions on cybernetics. 2019;50(11):4772-82.
77. Yoshida K, Yoshimoto J, Doya K. Sparse kernel canonical correlation analysis for discovery of nonlinear interactions in high-dimensional data. BMC bioinformatics. 2017;18(1):108.
78. Glaab E, Rauschenberger A, Banzi R, Gerardi C, Garcia P, Demotes J. Biomarker discovery studies for patient stratification using machine learning analysis of omics data: a scoping review. BMJ open. 2021;11(12):e053674.
79. Chen T, Tyagi S. Integrative computational epigenomics to build data-driven gene regulation hypotheses. GigaScience. 2020;9(6):giaa064.

 doi:10.20944/preprints202605.0428.v1

80. Ma Y, Sun Z, Zeng P, Zhang W, Lin Z. JSNMF enables effective and accurate integrative analysis of single-cell multiomics data. Briefings in bioinformatics. 2022;23(3):bbac105.
81. Uyar B, Savchyn T, Naghsh Nilchi A, Sarigun A, Wurmus R, Shaik MM, et al. Flexynesis: A deep learning toolkit for bulk multi-omics data integration for precision oncology and beyond. Nature Communications. 2025;16(1):8261.
82. Wang W, Zhang X, Dai D-Q. DeFusion: a denoised network regularization framework for multi-omics integration. Briefings in Bioinformatics. 2021;22(5):bbab057.
83. Ma T, Zhang A. Integrate multi-omics data with biological interaction networks using Multi-view Factorization AutoEncoder (MAE). BMC genomics. 2019;20(Suppl 11):944.
84. Rodrigues RR, Shulzhenko N, Morgun A. Transkingdom networks: a systems biology approach to identify causal members of host–microbiota interactions. Microbiome Analysis: Methods and Protocols: Springer; 2018. p. 227-42.
85. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics. 2008;9(1):559.
86. Glass K, Huttenhower C, Quackenbush J, Yuan G-C. Passing messages between biological networks to refine predicted interactions. PloS one. 2013;8(5):e64832.
87. Gao C, Yang C, Zhang L. A graph neural network-based spatial multi-omics data integration method for deciphering spatial domains. PLOS Computational Biology. 2025;21(9):e1013546.
88. Miao J, Li J, Xin J, Tu J, Ge M, Qi J, et al. MultiGATE: integrative analysis and regulatory inference in spatial multi-omics data via graph representation learning. Nature Communications. 2025;16(1):9403.
89. Kim D, Li R, Dudek SM, Ritchie MD. ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network. BioData mining. 2013;6(1):23.
90. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-87. e29.
91. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-cell multi-omic integration compares and contrasts features of brain cell identity. Cell. 2019;177(7):1873-87. e17.
92. Simidjievski N, Bodnar C, Tariq I, Scherer P, Andres Terre H, Shams Z, et al. Variational autoencoders for cancer data integration: design principles and computational practice. Frontiers in genetics. 2019;10:1205.
93. Acharya D, Mukhopadhyay A. A comprehensive review of machine learning techniques for multi-omics data integration: challenges and applications in precision oncology. Briefings in functional genomics. 2024;23(5):549-60.
94. Dihazi H, Asif AR, Beißbarth T, Bohrer R, Feussner K, Feussner I, et al. Integrative omics-from data to biology. Expert review of proteomics. 2018;15(6):463-6.
95. Papadaki E, Kakkos I, Vlamos P, Petropoulou O, Miloulis ST, Palamas S, et al. Recent web platforms for multi-omics integration unlocking biological complexity. Applied Sciences. 2024;15(1):329.
96. Zhang J, Che Y, Liu R, Wang Z, Liu W. Deep learning–driven multi-omics analysis: enhancing cancer diagnostics and therapeutics. Briefings in bioinformatics. 2025;26(4):bbaf440.
97. Ozaki Y, Broughton P, Abdollahi H, Valafar H, Blenda AV. Integrating omics data and AI for cancer diagnosis and prognosis. Cancers. 2024;16(13):2448.
98. López de Maturana E, Alonso L, Alarcón P, Martín-Antoniano IA, Pineda S, Piorno L, et al. Challenges in the integration of omics and non-omics data. Genes. 2019;10(3):238.
99. Mirza B, Wang W, Wang J, Choi H, Chung NC, Ping P. Machine learning and integrative analysis of biomedical big data. Genes. 2019;10(2):87.
100. Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. Frontiers in genetics. 2017;8:84.
101. Agamah FE, Bayjanov JR, Niehues A, Njoku KF, Skelton M, Mazandu GK, et al. Computational approaches for network-based integrative multi-omics analysis. Frontiers in Molecular Biosciences. 2022;9:967205.
102. Esteller M. Cancer epigenomics: DNA methylomes and histone-modification maps. Nature reviews genetics. 2007;8(4):286-98.
103. Kanwal R, Gupta S. Epigenetics and cancer. Journal of applied physiology. 2010.

 doi:10.20944/preprints202605.0428.v1

104. Chatterjee B, Majumder P, Chen C-C, Wang J-P, Su P-H, Lai H-C, et al. Hypoxia-induced genome-wide DNA demethylation by DNMT3A and EMT of cancer cells. Cellular & Molecular Biology Letters. 2025;30(1):95.
105. Plass C. Cancer epigenomics. Human molecular genetics. 2002;11(20):2479-88.
106. Teng P-C, Liang Y, Yarmishyn AA, Hsiao Y-J, Lin T-Y, Lin T-W, et al. RNA modifications and epigenetics in modulation of lung cancer and pulmonary diseases. International journal of molecular sciences. 2021;22(19):10592.
107. Bahado-Singh RO, Ibrahim A, Al-Wahab Z, Aydas B, Radhakrishna U, Yilmaz A, et al. Precision gynecologic oncology: circulating cell free DNA epigenomic analysis, artificial intelligence and the accurate detection of ovarian cancer. Scientific Reports. 2022;12(1):18625.
108. Kim YK, Kim WJ. Epigenetic markers as promising prognosticators for bladder cancer. International Journal of Urology. 2009;16(1):17-22.
109. Hudlikar R, Wang L, Wu R, Li S, Peter R, Shannar A, et al. Epigenetics/epigenomics and prevention of early stages of cancer by isothiocyanates. Cancer Prevention Research. 2021;14(2):151-64.
110. Mandal J, Aryal S, Manandhar I, Chakraborty S, Mei X, Yeoh BS, et al. Epigenetic histone β-hydroxybutyrylation contributes to renoprotection by β-hydroxybutyrate in the dahl rat. Hypertension. 2025;82(10):1729-42.
111. Basuroy T, Dreier M, Baum C, Blomquist T, Trumbly R, Filipp FV, et al. Epigenetic and pharmacological control of pigmentation via Bromodomain Protein 9 (BRD9). Pigment Cell & Melanoma Research. 2023;36(1):19-32.
112. Baca SC, Seo J-H, Davidsohn MP, Fortunato B, Semaan K, Sotudian S, et al. Liquid biopsy epigenomic profiling for cancer subtyping. Nature medicine. 2023;29(11):2737-41.
113. Moreno-Gonzalez M, Sierra I, Kind J. A hitchhiker's guide to single-cell epigenomics: Methods and applications for cancer research. International Journal of Cancer. 2026;158(2):291-304.
114. Moran S, Martinez-Cardús A, Boussios S, Esteller M. Precision medicine based on epigenomics: the paradigm of carcinoma of unknown primary. Nature reviews Clinical oncology. 2017;14(11):682-94.
115. Simo-Riudalbas L, Esteller M. Cancer genomics identifies disrupted epigenetic genes. Human genetics. 2014;133(6):713-25.
116. Tan WY, Nagabhyrava S, Ang-Olson O, Das P, Ladel L, Sailo B, et al. Translation of epigenetics in cell-free DNA liquid biopsy technology and precision oncology. Current Issues in Molecular Biology. 2024;46(7):6533-65.
117. Lomberk G, Dusetti N, Iovanna J, Urrutia R. Emerging epigenomic landscapes of pancreatic cancer in the era of precision medicine. Nature communications. 2019;10(1):3875.
118. Song D, Zhang Z, Zheng J, Zhang W, Cai J. 5-Hydroxymethylcytosine modifications in circulating cell-free DNA: frontiers of cancer detection, monitoring, and prognostic evaluation. Biomarker research. 2025;13(1):39.
119. Thomson JP, Meehan RR. The application of genome-wide 5-hydroxymethylcytosine studies in cancer research. Epigenomics. 2017;9(1):77-91.
120. Rodríguez-Paredes M, Esteller M. Cancer epigenetics reaches mainstream oncology. Nature medicine. 2011;17(3):330-9.
121. Herceg Z, Vaissière T. Epigenetic mechanisms and cancer: an interface between the environment and the genome. Epigenetics. 2011;6(7):804-19.
122. Hillyar C, Rallis KS, Varghese J, Hillyar CR. Advances in epigenetic cancer therapeutics. Cureus. 2020;12(11).
123. Krupinski J, Carrera C, Muiño E, Torres N, Al-Baradie R, Cullell N, et al. DNA methylation in stroke. Update of latest advances. Computational and Structural Biotechnology Journal. 2018;16:1-5.
124. Atzemian N, Mohammed S, Di Venanzio L, Gorica E, Costantino S, Ruschitzka F, et al. Cardiometabolic disease management: influences from epigenetics. Epigenomics. 2025;17(7):463-74.
125. Amenyah SD, Ward M, Strain J, McNulty H, Hughes CF, Dollin C, et al. Nutritional epigenomics and age-related disease. Current Developments in Nutrition. 2020;4(7):nzaa097.

 doi:10.20944/preprints202605.0428.v1

126. Lubin FD, Gupta S, Parrish RR, Grissom NM, Davis RL. Epigenetic mechanisms: critical contributors to long-term memory formation. The Neuroscientist. 2011;17(6):616-32.
127. Deogharia M, Gurha P. Epigenetic regulation of heart failure. Current opinion in cardiology. 2024;39(4):371-9.
128. O'Donnell KJ, Meaney MJ. Epigenetics, development, and psychopathology. Annual Review of Clinical Psychology. 2020;16(1):327-50.
129. Rönn T, Perfilyev A, Oskolkov N, Ling C. Predicting type 2 diabetes via machine learning integration of multiple omics from human pancreatic islets. Scientific Reports. 2024;14(1):14637.
130. Hu C, Li Y, Li L, Zhang N, Zheng X. BS-clock, advancing epigenetic age prediction with high-resolution DNA methylation bisulfite sequencing data. Bioinformatics. 2024;40(11):btae656.
131. Wu Z, Herring D, Spill F, Andrews J, editors. iTARGET: Interpretable Tailored Age Regression for Grouped Epigenetic Traits. 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2024: IEEE.
132. Xu Z. DNA methylation-based health predictors: advances, applications, and perspectives. Epigenomics. 2025;17(15):1083-90.
133. Zhou W, Johnson BK, Morrison J, Beddows I, Eapen J, Katsman E, et al. BISCUIT: an efficient, standards-compliant tool suite for simultaneous genetic and epigenetic inference in bulk and single-cell studies. Nucleic Acids Research. 2024;52(6):e32-e.
134. Mullen DJ, Wu Z, Nelson-Moore E, Cao H, Han L, Offringa IA, et al. Tracing regulatory element networks using epigenetic traits to identify key transcription factors: TENET R/Bioconductor package. Bioinformatics. 2025;41(8):btaf435.
135. Couckuyt A, Seurinck R, Emmaneel A, Quintelier K, Novak D, Van Gassen S, et al. Challenges in translational machine learning. Human Genetics. 2022;141(9):1451-66.
136. Hsu C-Y, Askar S, Alshkarchy SS, Nayak PP, Attabi KA, Khan MA, et al. AI-driven multi-omics integration in precision oncology: bridging the data deluge to clinical decisions. Clinical and experimental medicine. 2026;26(1):29.
137. Li N, Song K, Chen H, Dai M. Advance and challenge of DNA methylation as cancer biomarkers for risk stratification, screening and early detection. Journal of the National Cancer Center. 2025;5(2):108-12.
138. Bonora G, Rubbi L, Morselli M, Ma F, Chronis C, Plath K, et al. DNA methylation estimation using methylation-sensitive restriction enzyme bisulfite sequencing (MREBS). PLoS One. 2019;14(4):e0214368.
139. Derks S, Lentjes MH, Hellebrekers DM, De Bruïne AP, Herman JG, Van Engeland M. Methylation-specific PCR unraveled. Analytical Cellular Pathology. 2004;26(5-6):291-9.
140. Shaw RJ, Akufo-Tetteh EK, Risk JM, Field JK, Liloglou T. Methylation enrichment pyrosequencing: combining the specificity of MSP with validation by pyrosequencing. Nucleic acids research. 2006;34(11):e78-e.
141. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome biology. 2012;13(10):R87.
142. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome biology. 2008;9(9):R137.
143. Stark R, Brown G. DiffBind: differential binding analysis of ChIP-Seq peak data. R package version. 2011;100(4.3):2-21.
144. Mao Z, Li T, Zhao H, Qin Y, Wang X, Kang Y. Identification of epigenetic interactions between microRNA and DNA methylation associated with polycystic ovarian syndrome. Journal of Human Genetics. 2021;66(2):123-37.
145. Chauvel C, Novoloaca A, Veyre P, Reynier F, Becker J. Evaluation of integrative clustering methods for the analysis of multi-omics data. Briefings in bioinformatics. 2020;21(2):541-52.
146. Zhang S, Lv J, Zhang J, Fan Z, Gu B, Fan B, et al. Benchmarking multi-omics integrative clustering methods for subtype identification in colorectal cancer. Computer Methods and Programs in Biomedicine. 2025;261:108603.

 doi:10.20944/preprints202605.0428.v1

147. Cao Y, Zhao X, Tang S, Jiang Q, Li S, Li S, et al. scButterfly: a versatile single-cell cross-modality translation method via dual-aligned variational autoencoders. Nature Communications. 2024;15(1):2973.
148. Xiao C, Chen Y, Meng Q, Wei L, Zhang X. Benchmarking multi-omics integration algorithms across single-cell RNA and ATAC data. Briefings in Bioinformatics. 2024;25(2):bbae095.
149. Dozmorov MG. Epigenomic annotation-based interpretation of genomic data: from enrichment analysis to machine learning. Bioinformatics. 2017;33(20):3323-30.
150. Taudt A, Colomé-Tatché M, Johannes F. Genetic sources of population epigenomic variation. Nature Reviews Genetics. 2016;17(6):319-32.
151. Santaló J, Berdasco M. Ethical implications of epigenetics in the era of personalized medicine. Clinical epigenetics. 2022;14(1):44.
152. Yassi M, Chatterjee A, Parry M. Application of deep learning in cancer epigenetics through DNA methylation analysis. Briefings in bioinformatics. 2023;24(6):bbad411.
153. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nature Reviews Genetics. 2015;16(6):321-32.
154. Tahir M, Norouzi M, Khan SS, Davie JR, Yamanaka S, Ashraf A. Artificial intelligence and deep learning algorithms for epigenetic sequence analysis: A review for epigeneticists and AI experts. Computers in Biology and Medicine. 2024;183:109302.
155. Sahoo K, Lingasamy P, Khatun M, Sudhakaran SL, Salumets A, Sundararajan V, et al. Artificial Intelligence in cancer epigenomics: a review on advances in pan-cancer detection and precision medicine. Epigenetics & chromatin. 2025;18(1):35.