applied sciences

MDPI

Review

Epigenomic Biomarker Discovery from Biomedical Literature: AI and Text Mining Toward Health Monitoring Frameworks

Ji-Hye Oh [1], Hee-Jo Nam [1], Soo Hyun Seo [2] and Hyun-Seok Park [1,3,4,*]

[1] Department of Computer Science & Engineering, ELTEC College of Engineering, Ewha Womans University, Seoul 03760, Republic of Korea; ohji84200@ewhain.net (J.-H.O.); heejonam@ewhain.net (H.-J.N.)
[2] Department of Laboratory Medicine, Seoul National University Bundang Hospital, 82 Gumi-ro 173beon-gil, Bundang-gu, Seongnam-si 13620, Gyeonggi-do, Republic of Korea; shseo@snubh.org
[3] Department of Computer Science & Engineering, College of AI, Ewha Womans University, Seoul 03760, Republic of Korea
[4] Department of Computational Medicine, Ewha Womans University, Seoul 03760, Republic of Korea
* Correspondence: neo@ewha.ac.kr; Tel.: +82-232773513

Abstract

Epigenomic regulation, particularly DNA methylation, plays a critical role in gene expression control and has emerged as an important source of biomarkers for disease diagnosis, risk prediction, and longitudinal health monitoring. As high-throughput sequencing technologies have expanded, epigenomic research has rapidly grown, producing a large and complex body of biomedical literature. This review presents an AI-driven literature-level analysis aimed at uncovering structural patterns and research trends related to epigenomic biomarker discovery. Using a large corpus of full-text articles collected from PubMed and PubMed Central, we applied text mining techniques including keyword frequency analysis, document-level co-occurrence analysis, topic clustering, contextual concordance analysis, and temporal trend analysis. Rather than evaluating individual experiments, this approach examines the broader research landscape to identify recurring conceptual structures and methodological patterns. The analysis reveals that epigenomic biomarker research is organized into several interconnected domains, including disease-focused epigenomics, chromatin regulation studies, transcriptomic integration research, and cancer-related epigenomic investigations. The rapid growth of publications since 2010 further reflects the increasing importance of high-throughput epigenomic profiling and biomarker-driven research. These findings demonstrate that AI-driven literature mining provides a scalable framework for uncovering epigenomic biomarker knowledge and translating it toward AI-enabled health monitoring systems. Such approaches may support biomarker prioritization, early disease detection, and data-driven health monitoring within precision health environments.

Keywords: epigenomics; DNA methylation; text mining; literature-level analysis; PubMed; genetic loci; chromatin; experimental methods; literature-level patterns

Check for updates

Academic Editor: Keun Ho Ryu

Received: 31 March 2026
Revised: 30 April 2026
Accepted: 30 April 2026
Published: 8 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Epigenomic regulation plays a critical role in human disease by modulating gene expression without altering the underlying DNA sequence. In addition to genetic mutations, mechanisms such as DNA methylation, histone modifications, and chromatin organization contribute substantially to disease development, particularly in cancer and other complex

disorders. Among these, DNA methylation has been widely studied as a stable and disease-associated signal, frequently linked to transcriptional silencing of tumor suppressor genes. Importantly, epigenomic alterations are dynamic and context-dependent, making them promising candidates for biomarkers in disease detection, prognosis, and monitoring.

The rapid advancement of high-throughput sequencing technologies has led to a dramatic expansion of epigenomic research across diverse biological systems. As a result, the volume of related literature has grown extensively, creating challenges in synthesizing accumulated knowledge and identifying consistent patterns across studies. While many existing reviews focus on specific diseases or molecular mechanisms, fewer studies have examined the broader structural organization of epigenomic research at the literature level.

Recent developments in artificial intelligence (AI) and text mining provide new opportunities to address this challenge. Literature-level analysis enables the systematic extraction of recurring concepts, co-occurring research themes, and temporal trends from large-scale document collections. Rather than focusing on individual experimental findings, such approaches reveal higher-level patterns in how epigenomic biomarkers are studied, reported, and conceptualized across the scientific literature.

In this review, we survey the biomedical literature on epigenomic regulation, with a focus on DNA methylation and related regulatory features, using a literature-level AI and text mining framework to identify recurring patterns and biomarker-relevant structures. By organizing structural relationships within the literature, this review aims to characterize biomarker-relevant signals and to discuss how such literature-derived insights may inform future AI-enabled health monitoring frameworks in which epigenomic biomarkers serve as a molecular layer within multimodal systems.

For clarity, key technical terms used throughout this review are summarized in Table A1 in Appendix A.

2. Epigenomic Biomarkers and Chromatin Variability in Human Health

Epigenomic regulation provides a multi-layered framework for understanding gene expression control and its disruption in human disease. Unlike static genetic mutations, epigenomic features are dynamic, reversible, and context-dependent, making them particularly suitable as biomarkers for disease detection, prognosis, and longitudinal health monitoring. In this review, we consider a set of core epigenomic signals—including DNA methylation, chromatin variability, genetic variation (SNPs), histone modifications, and non-coding RNAs—as a unified signal space for discussing biomarker discovery at the literature level. Within this framework, epigenomic biomarkers are interpreted not as isolated molecular entities, but as interconnected components of a broader regulatory system.

2.1. DNA Methylation as a Core Biomarker Layer

DNA methylation is one of the most extensively studied epigenomic mechanisms and has been widely investigated as a source of biomarkers for disease diagnosis, prognostic prediction, and therapeutic monitoring [1]. Aberrant methylation patterns, particularly hypermethylation in promoter regions of tumor suppressor genes, are strongly associated with transcriptional silencing and tumorigenesis [2]. Beyond oncology, methylation alterations are also implicated in metabolic, cardiovascular, and neurological disorders, highlighting their broad applicability.

Importantly, DNA methylation is dynamically regulated and reflects environmental exposures, aging processes, and disease progression. These properties make it well suited for longitudinal monitoring applications. Advances in high-throughput sequencing and single-cell epigenomics have further enabled the identification of disease-associated methylation signatures across diverse biological contexts.

2.2. Epigenetic Aging and Monitoring-Relevant Signals

In addition to disease-specific biomarkers, DNA methylation serves as a quantitative indicator of biological aging. Epigenetic clock models estimate biological age based on methylation signatures and capture cumulative effects of environmental and physiological factors [3,4]. Because methylation changes can precede clinical symptoms, these biomarkers are particularly relevant for early risk stratification and preventive health monitoring.

Such properties position methylation-based biomarkers as a foundational molecular layer in emerging digital health systems, where long-term molecular signals can complement high-frequency physiological monitoring data.

2.3. Chromatin Variability and Regulatory Dynamics

Chromatin structure provides an additional layer of epigenomic regulation that reflects dynamic changes in gene accessibility across cell types and conditions. Large-scale epigenomic projects have demonstrated that chromatin states vary substantially across biological contexts, and this variability can be quantified using chromatin state annotations [5,6].

Chromatin state variability captures the degree to which a genomic locus exhibits different regulatory states across multiple epigenomes. Regions with high variability often correspond to regulatory elements such as enhancers, whereas more stable regions tend to reflect structural or constitutive genomic functions [7,8]. These patterns suggest that chromatin variability can serve as an indicator of regulatory importance and a potential criterion for prioritizing candidate biomarkers.

Recent machine learning approaches further demonstrate that chromatin variability can be partially predicted from DNA sequence features, indicating that sequence-encoded regulatory information contributes to epigenomic plasticity [8]. This provides a conceptual link between genomic sequence, regulatory activity, and epigenomic variability.

2.4. Genetic Variation and SNP–Epigenome Interactions

Genetic variation, particularly single nucleotide polymorphisms (SNPs), represents an additional layer influencing epigenomic regulation. Disease-associated SNPs identified through genome-wide association studies (GWAS) are frequently located in non-coding regions, suggesting regulatory roles mediated through epigenomic mechanisms.

Emerging evidence indicates that genetic variants can affect DNA methylation patterns, chromatin accessibility, and transcription factor binding. However, the relationship between SNPs and chromatin variability remains incompletely characterized. Understanding how genetic variation constrains or modulates epigenomic dynamics is therefore essential for integrating genetic and epigenetic biomarkers within a unified framework.

2.5. Histone Modifications and Chromatin State Definition

Histone modifications, including acetylation and methylation, play a central role in defining chromatin states and regulating gene expression. Specific histone marks, such as H3K27ac and H3K4me3, are associated with active regulatory regions, while others correspond to repressive chromatin states.

These modifications function in coordination with DNA methylation and chromatin accessibility to shape the regulatory landscape of the genome. As such, histone modification patterns provide complementary information for identifying functional regulatory elements and refining biomarker interpretation.

2.6. Non-Coding RNAs as Regulatory Biomarkers

Non-coding RNAs, including microRNAs (miRNAs) and long non-coding RNAs (lncRNAs), represent an additional regulatory layer influencing gene expression. These

molecules often exhibit disease-specific expression patterns and have been widely investigated as circulating biomarkers in clinical settings.

Importantly, non-coding RNAs frequently interact with epigenomic mechanisms, including DNA methylation and chromatin regulation, forming integrated regulatory networks. Their inclusion expands the biomarker space beyond DNA-level modifications to encompass transcriptional and post-transcriptional regulation.

2.7. Toward a Unified Epigenomic Biomarker Space

Together, these components define a multi-layered epigenomic signal space for biomarker discovery. DNA methylation provides a stable and measurable molecular baseline, chromatin variability captures regulatory dynamics, genetic variation introduces inherited constraints, while histone modifications and non-coding RNAs add complementary regulatory information.

This integrated perspective enables the systematic exploration of biomarker candidates at the literature level. In the following sections, we describe how large-scale text mining can be used to extract and organize these signals into structured, biomarker-relevant patterns.

3. Literature Mining Framework for Epigenomic Biomarker Discovery

Recent advances in biomedical text mining have substantially transformed the landscape of knowledge extraction for biomarker discovery [9,10]. Early approaches primarily relied on rule-based systems and statistical co-occurrence analysis to identify gene–disease associations and functional relationships [11–13]. However, the field has rapidly evolved with the introduction of deep learning and transformer-based language models specifically tailored to biomedical corpora. Models such as BioBERT and PubMedBERT have demonstrated strong performance in named entity recognition, relation extraction, and semantic similarity tasks, enabling more accurate identification of biomarker-related entities and interactions within large-scale literature datasets [14–16].

In parallel, large-scale resources such as PubTator and the Comparative Toxicogenomics Database (CTD) have facilitated structured knowledge integration by linking genes, diseases, chemicals, and pathways extracted from literature [17,18]. Recent studies have further extended these approaches to construct literature-derived knowledge graphs and multimodal representations that integrate text-mined features with genomic and epigenomic datasets [19,20].

Despite these advances, most prior work has focused on improving extraction accuracy at the entity or relation level. In contrast, relatively fewer studies have explored literature-level structural patterns and their potential role in guiding biomarker prioritization and translational health monitoring frameworks [9,13]. This gap highlights the need for complementary approaches that emphasize large-scale conceptual organization and knowledge synthesis across the biomedical literature. Accordingly, the present study adopts a literature-level analytical perspective, focusing on structural and conceptual patterns rather than model-level performance optimization.

To support this review, we organized a literature-level framework for surveying epigenomic biomarker research at scale. In this section, we describe the rationale for using AI and text mining approaches, the construction of the literature corpus, the query design used for structured retrieval, and the analytical components used to identify recurring methodological, thematic, contextual, and temporal patterns across the reviewed literature.

3.1. AI and Literature-Level Analysis in Epigenomic Research

The rapid expansion of biomedical data has accelerated the adoption of artificial intelligence (AI) and machine learning approaches for biomarker discovery and health

prediction [21]. In epigenomic research, AI methods have been widely applied to DNA methylation data for disease classification, biological age prediction, and early disease detection. Deep learning approaches have enabled the identification of tumor-specific methylation signatures across multiple cancer types, while epigenetic clock models have improved the estimation of biological age and disease risk.

Despite these advances, most studies focus on experimental datasets or predictive modeling, with limited attention to the broader structure of epigenomic research at the literature level. Understanding how biomarker-related knowledge is distributed, organized, and evolves across studies requires systematic analysis of large-scale scientific literature rather than isolated experimental findings.

Biomedical literature mining has emerged as a scalable approach to address this challenge. Research in this area has evolved across both biomedical informatics and computer science communities, emphasizing interpretability and domain knowledge integration on one hand, and scalability and predictive performance on the other [22]. Advances in distributed computing and machine learning have further enabled large-scale text analysis, making it feasible to process thousands of biomedical articles efficiently.

Importantly, full-text mining provides substantially richer information than abstract-based approaches. Full articles contain detailed experimental contexts, methodological descriptions, and biological relationships that are often absent from abstracts [21]. These characteristics make full-text corpora particularly suitable for extracting biomarker-relevant knowledge and reconstructing literature-level structures in epigenomic research.

3.2. Data Sources and Corpus Construction

For this review, the literature corpus was compiled from epigenomic and DNA methylation-related research articles retrieved from PubMed and PubMed Central (PMC). To enable large-scale and reproducible analysis, only open-access articles with available full text were included [23,24].

Literature retrieval was performed using core keywords and related expressions associated with epigenetics and DNA methylation. No restrictions were imposed on publication year, disease type, or experimental design to ensure broad coverage of epigenomic biomarker research. Full-text materials—including titles, abstracts, and article bodies—were collected using PubMed Central identifiers (PMCIDs) and stored in both plain-text and structured XML formats.

The study utilized the BioC data format, which provides a standardized XML/JSON representation of biomedical literature and supports efficient computational processing. BioC-based PMC collections enable consistent annotation and scalable corpus management. Using this strategy, a total of 6152 full-text articles published before 27 January 2024 were included in the final corpus.

3.3. Query Design for Epigenomic Biomarker Retrieval

To construct a comprehensive and structured corpus, query patterns were designed to capture a wide range of epigenomic biomarker studies. The retrieval strategy integrates core epigenetic concepts—such as DNA methylation, histone modification, and chromatin regulation—with biomarker-related terminology [21].

A two-tier query design was adopted to balance coverage and specificity:

- Baseline queries were constructed to maximize recall by broadly capturing epigenetic biomarker literature across diverse contexts.
- Extended queries incorporated constraints such as disease categories, mechanistic descriptors (e.g., hypermethylation), clinical cohort indicators, and machine learning-related terminology to improve specificity.

Table 1 presents structured baseline query patterns designed for the high-recall retrieval of epigenetic biomarker-related literature. These queries combine core epigenetic concepts with biomarker-related expressions to capture diverse research contexts, including diagnostic, prognostic, and multi-omics studies.

Table 1. Exemplary query patterns for epigenetic biomarker retrieval in large-scale text mining. This table presents baseline query patterns designed to maximize recall in retrieving epigenetic biomarker-related literature. The queries combine core epigenetic concepts, including DNA methylation, histone modification, and chromatin accessibility, with biomarker-related terms to capture diverse research contexts such as diagnostic, prognostic, and multi-omics studies.

Objective	Query Pattern	Application Context	Notes
General biomarker screening	(epigenetic OR "DNA methylation") AND biomarker	Broad initial corpus retrieval	High recall
Diagnostic biomarker	(DNA methylation OR histone) AND "diagnostic biomarker"	Early detection marker identification	Strong clinical relevance
Prognostic biomarker	(epigenetic OR methylation) AND prognostic AND survival	Survival analysis-related studies	Frequently linked to TCGA datasets
Signature discovery	(epigenetic AND signature) AND cancer	Multi-gene biomarker identification	Often includes machine learning studies
Specific histone modification	(H3K27ac OR H3K4me3) AND biomarker	Histone modification-based markers	Enhancer-associated biomarkers
CpG marker	("CpG island methylation") AND biomarker	Promoter silencing markers	Tumor suppressor gene studies
Chromatin accessibility biomarker	(ATAC-seq OR "chromatin accessibility") AND biomarker	Regulatory biomarker discovery	Increasing in recent studies
Multi-omics biomarker	(methylation AND expression) AND biomarker	Integrative inverse-correlation markers	Multi-omics integration studies
Liquid biopsy biomarker	("circulating DNA methylation") AND biomarker	Non-invasive diagnostic biomarkers	High translational value
Drug response biomarker	(epigenetic AND "drug response") AND biomarker	Therapy response markers	Precision medicine applications
Aging biomarker	("epigenetic clock" OR "methylation age") AND biomarker	Aging and biological age studies	Includes Horvath clock-based models
Neuro-epigenetic biomarker	("DNA methylation" AND brain) AND biomarker	Neurodegenerative disease markers	Alzheimer's disease research
Immune-related epigenetic biomarker	(epigenetic AND immune) AND biomarker	Immunotherapy-related biomarkers	Checkpoint response studies
Enhancer biomarker	("super-enhancer" OR enhancer) AND epigenetic biomarker	Transcriptional regulation markers	Cancer subtype characterization
Transcription factor regulation biomarker	("ChIP-seq" AND "transcription factor") AND biomarker	Regulatory network biomarkers	Often linked to ENCODE data

Table 2 introduces precision-enhancing extended query patterns aimed at reducing retrieval noise and increasing thematic specificity. These strategies incorporate disease restrictions, mechanistic descriptors, clinical indicators, and validation-focused criteria.

Table 2. Precision-enhancing extended query patterns for epigenomic biomarker retrieval. This table summarizes extended query strategies aimed at improving retrieval specificity by incorporating constraints such as disease categories, mechanistic descriptors (e.g., hypermethylation), clinical cohort indicators, and machine learning-related terminology.

Strategy	Query Pattern	Intended Effect
Disease restriction	(“epigenetic biomarker”) AND (cancer OR tumor)	Reduces domain noise by restricting disease scope
Mechanism emphasis	(hypermethylation OR hypomethylation) AND biomarker	Focuses on mechanistically characterized biomarkers
Clinical relevance	(“epigenetic biomarker”) AND cohort	Prioritizes patient-based or population-level studies
Machine learning-based discovery	(epigenetic AND “machine learning”) AND biomarker	Identifies computationally derived biomarker signatures
Validation-focused studies	(“epigenetic biomarker”) AND validation	Retrieves studies emphasizing reproducibility and independent validation

Together, this two-tier query design enables scalable yet structured corpus construction for literature-level epigenomic biomarker analysis.

3.4. Text Mining and Analytical Pipeline

To identify structural patterns in epigenomic literature, multiple complementary text mining techniques were applied.

First, keyword frequency and co-occurrence analyses were used to identify frequently occurring biological concepts, genomic elements, and experimental approaches. These methods reveal dominant research themes and conceptual associations across the literature.

Second, topic modeling was applied to identify latent thematic structures within the corpus, enabling the classification of studies into major research domains [22].

Third, concordance-based contextual analysis was used to examine how specific epigenomic concepts—such as genomic loci, chromatin states, and methylation patterns—are used across different narrative contexts.

Finally, temporal trend analysis was conducted to investigate the evolution of research themes and methodological practices over time.

These analytical components form an integrated analytical framework. The overall literature selection process is illustrated in Figure 1, while the analytical workflow is conceptually described in this section. This figure illustrates the overall workflow for integrating large-scale epigenomic data with genetic and regulatory annotations. For each genomic unit, chromatin states across all epigenomes are aggregated to compute a chromatin variability count, defined as the number of distinct states observed. Subsequently, genetic variation data, including SNP annotations from dbSNP and ClinVar, are mapped onto the same genomic coordinates, followed by the integration of regulatory features such as promoters and enhancers.

All processed information is consolidated into a unified database, which provides a genome-wide representation of chromatin variability, dominant states, and associated genetic and regulatory annotations. This integrated framework enables systematic investigation of the relationship between DNA sequence features, epigenomic stability, and functional genomic elements.

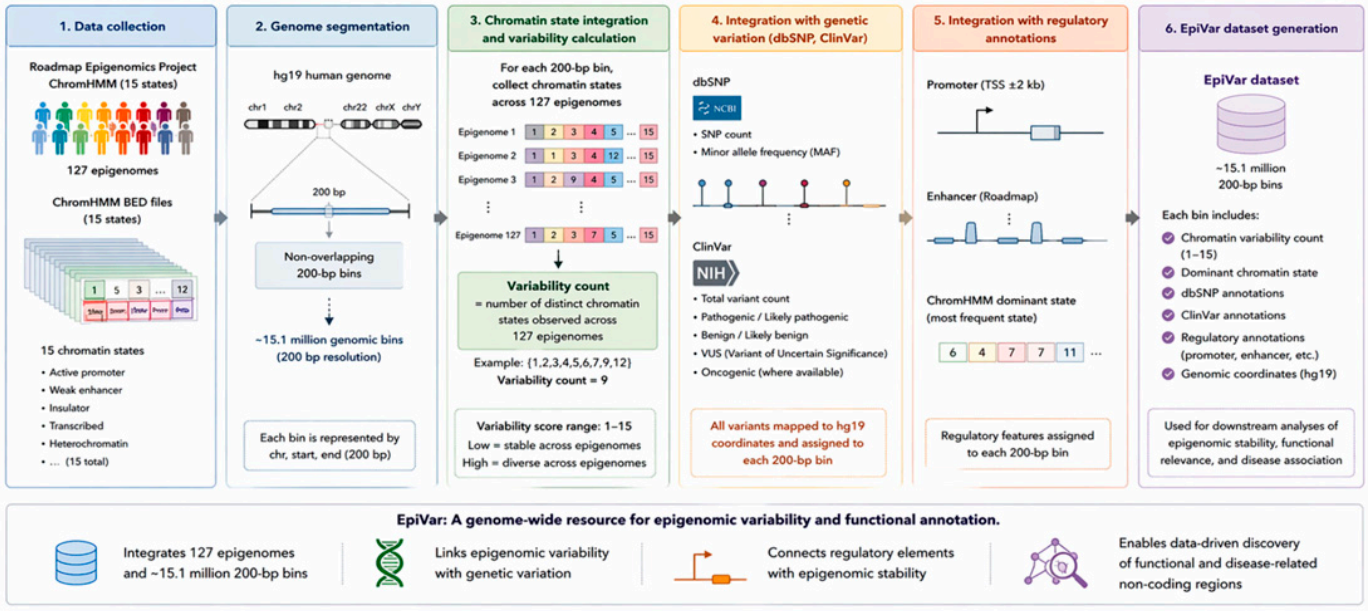


Figure 1. Conceptual workflow of literature retrieval and corpus construction for epigenomic biomarker analysis. The diagram illustrates the overall process of collecting full-text articles, including duplicate removal, title/abstract screening, and full-text eligibility assessment. This workflow supports the construction of a large-scale literature corpus used for downstream text mining analyses, including keyword frequency analysis, co-occurrence analysis, topic modeling, concordance-based contextual analysis, and temporal trend analysis. The figure is intended to provide a conceptual overview of the literature acquisition and preprocessing steps, rather than to represent a fully standardized systematic review protocol.

4. Extracted Patterns and Epigenomic Signals

In this section, we review the major patterns identified from the literature corpus and discuss their relevance to epigenomic biomarker research. The following subsections summarize the overall corpus characteristics, recurring methodological configurations, thematic domains, contextual regularities, and temporal trends observed across the literature. Taken together, these findings provide a structured literature-level perspective on biomarker-relevant signals and their broader translational implications.

4.1. Corpus Characteristics and Analytical Scope

The corpus examined in this review consists of a large collection of epigenomic and DNA methylation-related full-text articles spanning multiple decades. To capture the broad landscape of epigenomic research, the dataset was constructed without restrictions on disease type, experimental design, or biological system [25]. This inclusive strategy enables the comprehensive representation of diverse research domains and provides a robust foundation for literature-level analysis.

Each document includes complete textual components, including titles, abstracts, and full-text bodies, allowing detailed analysis of methodological descriptions and biological contexts [26]. Preprocessing procedures were applied to normalize textual content and metadata, ensuring consistency across heterogeneous publications, and enable large-scale computational analysis.

The resulting corpus contains thousands of full-text articles and tens of millions of words, highlighting the limitations of traditional manual review approaches. The scale of the dataset enables the detection of stable conceptual patterns, thematic clusters, and long-term research trends across the epigenomic literature.

The final corpus included 6152 full-text articles and 62,360,028 words in total, with an average of 10,136.5 words per article. These corpus statistics highlight the large-scale nature of the literature surveyed in this review and underscore the need for computational approaches to synthesize recurring patterns across epigenomic biomarker research.

4.2. Methodological Patterns in Epigenomic Research

Keyword and co-occurrence analyses reveal that epigenomic research is characterized by the widespread use of high-throughput sequencing and profiling techniques, including RNA-seq, ChIP-seq, DNase-seq, ATAC-seq, RRBS, and WGBS [27–30]. These technologies are frequently applied in recurring combinations depending on analytical objectives and biological contexts.

Rather than being used independently, experimental methods form consistent combinatorial patterns, reflecting standardized research design strategies across the literature. For example, RNA-seq often co-occurs with ChIP-seq in studies investigating transcriptional regulation, while chromatin accessibility assays are frequently paired with DNA methylation profiling techniques.

These recurring methodological configurations suggest the existence of stable experimental frameworks within epigenomic research. From a computational perspective, such regularities indicate that biomarker discovery pipelines can leverage standardized multi-omics integration strategies derived from literature-level evidence.

4.3. Thematic Structure of Epigenomic Research

Topic modeling reveals that the epigenomic literature is organized into multiple parallel research domains rather than a single dominant trajectory. Major thematic clusters include disease-centered studies, chromatin regulation and transcriptional control, integrative epigenomic–transcriptomic analyses, and cancer-focused research.

As summarized in Table 3, these clusters—derived from topic modeling of the literature corpus (e.g., LDA-based grouping with predefined topic numbers and standard preprocessing steps)—represent interconnected yet distinct research domains that collectively define the structure of epigenomic research. Importantly, these clusters are intended to capture high-level thematic trends and organizational patterns in the literature, rather than to provide fully optimized or quantitatively validated topic classifications. Rather than evolving linearly, the field has expanded through the parallel development of multiple complementary research areas.

This multi-domain structure provides a conceptual framework for organizing biomarker evidence, enabling the mapping of candidate biomarkers to specific biological contexts and research themes.

While conventional text mining methods provide interpretable insights, recent advances in transformer-based models enable contextual semantic representation of biomedical literature. An additional transformer-based semantic validation was conducted. This analysis was designed to examine whether the thematic structures identified through conventional text mining methods could also be observed in a contextual semantic representation of the corpus.

In Figure 2, each point represents an individual full-text article, and colors indicate semantic clusters identified through K-means clustering. PubMedBERT-based embeddings were reduced to two dimensions using UMAP. The visualization shows that the corpus forms structured semantic regions rather than a random distribution, supporting the presence of meaningful thematic organization in the epigenomic literature.

Table 3. Topic cluster composition and representative keywords identified through topic modeling. Topics were extracted using Latent Dirichlet Allocation (LDA). To validate cluster reliability, additional clustering analysis was performed using TF-IDF-based document representations and K-means clustering (Scikit-learn). TF-IDF parameters: max_features = 5000, min_df = 5, max_df = 0.85, stop-word removal applied. K-means parameters: n_clusters = 2–12, random_state = 42, n_init = 20, max_iter = 300. The optimal cluster number was determined using both the Elbow Method (WCSS) and Silhouette score, which consistently indicated stable clustering around k = 7. The LDA model achieved a Silhouette score of 0.504 and a coherence score (Cv) of 0.468, supporting both structural validity and semantic consistency of the extracted topics.

Cluster	Number of Groups	Cluster Words
Cluster 0	427	Plant, genome, assembly, species, sequence, chromosome
Cluster 1	1513	Patients, disease, variants, minimal, children
Cluster 2	620	Methylation, DNA, CpG, methylated, epigenetic, site, ages
Cluster 3	1119	Mice, RNA, mRNA
Cluster 4	824	Cancer, tumor, patients, immune, breast, survival, breast
Cluster 5	901	Chromatin, enhancer, peak, genome, transcription, DNA, binding, accessibility
Cluster 6	733	RNA, cancer, DNA, patients, disease

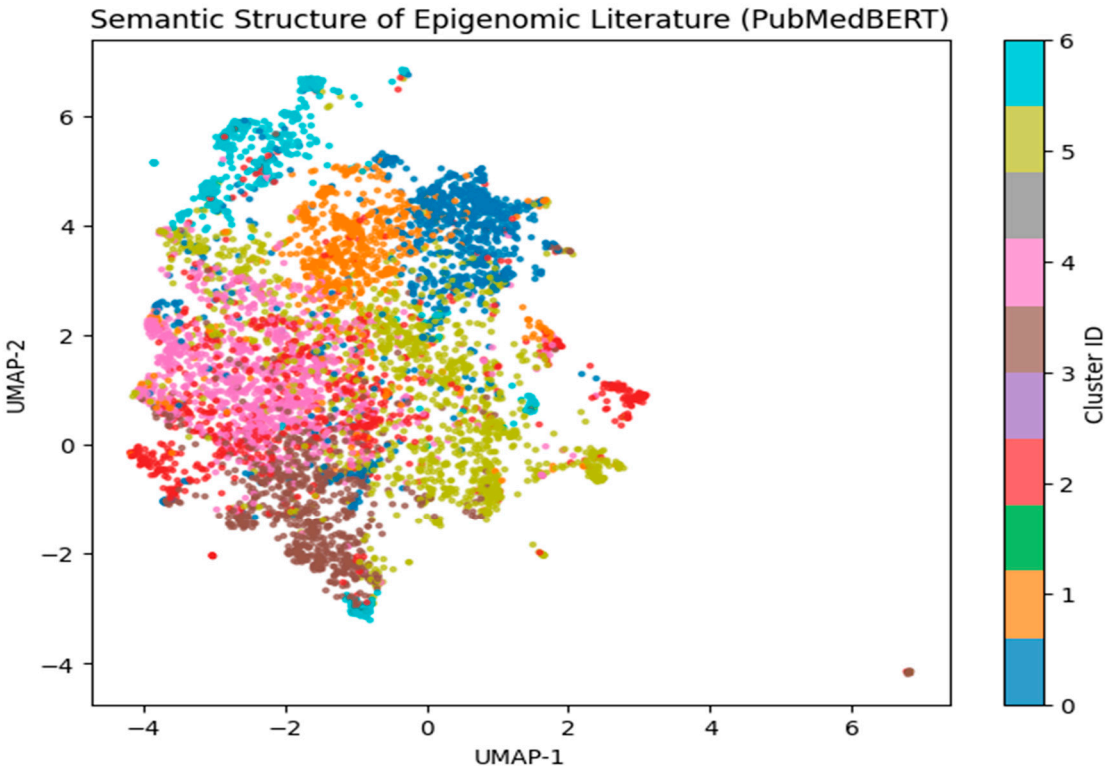


Figure 2. Semantic structure of epigenomic literature based on PubMedBERT embeddings. PubMedBERT-based sentence-transformer embeddings were generated from truncated full-text documents to capture contextual semantic relationships. The resulting high-dimensional representations were reduced to two dimensions using UMAP, followed by K-means clustering (n = 7) aligned with the LDA-derived topic structure. Each document is visualized in a 2D semantic space with cluster-based color coding. Representative documents and TF-IDF keywords were used to support cluster interpretation. This analysis is included as a complementary semantic validation step to assess the robustness of literature-derived patterns, rather than as a primary modeling approach. Colors indicate PubMedBERT-derived semantic clusters: **blue** (cancer epigenomics), **orange** (RNA epigenetic regulation), **green** (gene expression and RNA processing), **red** (developmental epigenetics), **purple** (bioinformatics resources), **brown** (gene regulation and signaling pathways), and **yellow** (global epigenetic mechanisms across systems).

4.4. Contextual Stability and Conceptual Consistency

Concordance-based analysis demonstrates that epigenomic terminology is embedded within highly consistent narrative structures across the literature. Core concepts, including genomic loci, chromatin states, and methylation-related processes, appear within recurring contextual patterns rather than randomly distributed textual fragments, as in Table 4 and Figure 3. Importantly, these observations are intended to provide a minimal, literature-level trend guideline rather than a fully optimized or exhaustive analytical characterization of epigenomic discourse.

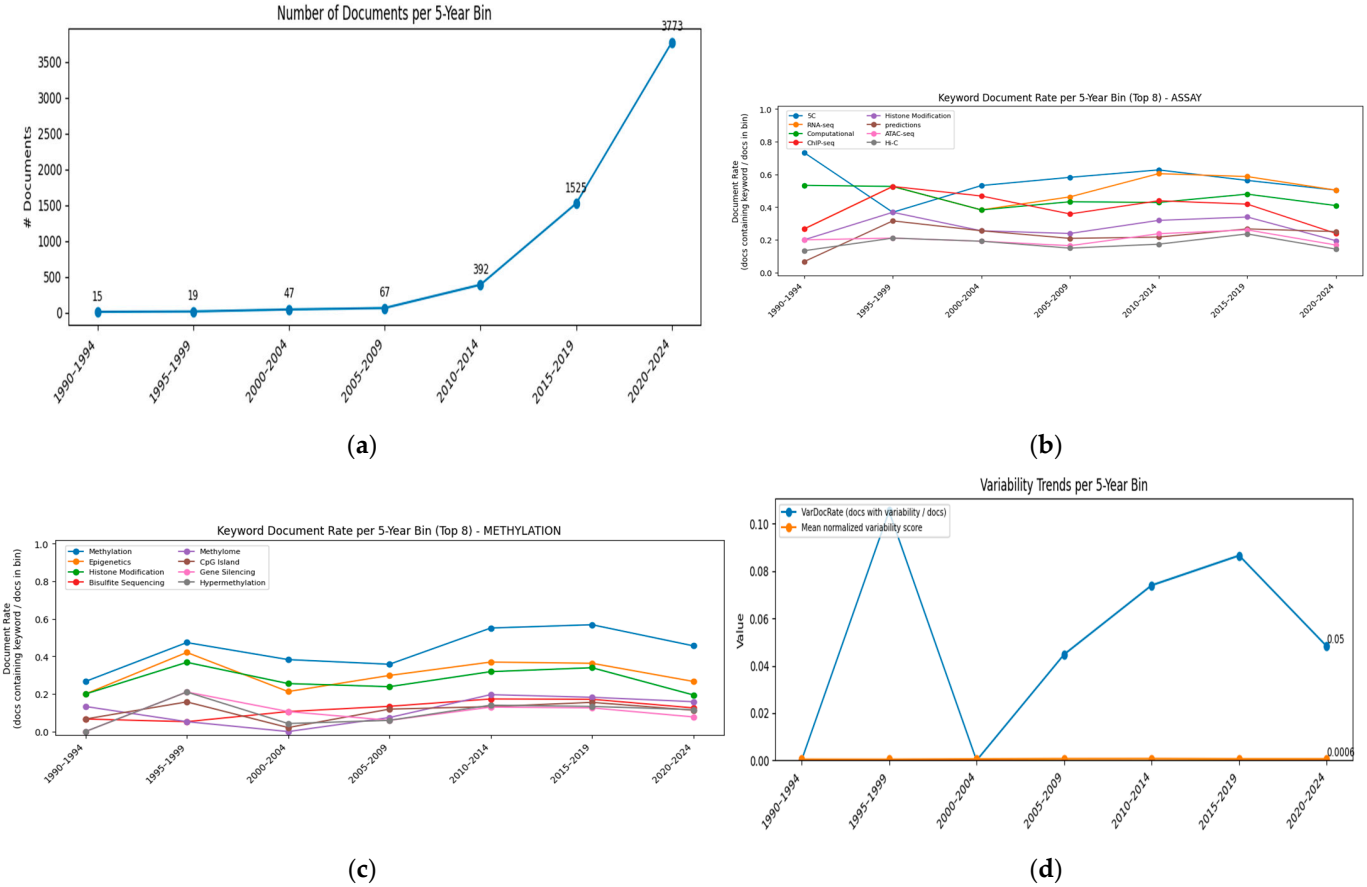


Figure 3. Temporal trends in epigenomics and DNA methylation research (1990–2024). (**a**) Publication counts aggregated in five-year intervals (BIN_SIZE = 5) within the range 1990–2025, including only bins with ≥10 documents (MIN_DOCS_PER_BIN = 10). (**b**) Assay-related keyword trends measured as document-level occurrence rates. The top eight keywords (TOP_K_ASSAY = 8) were selected based on mean frequency across bins and include sequencing and chromatin profiling techniques (e.g., RNA-seq, ChIP-seq, ATAC-seq, WGBS, Hi-C). (**c**) Methylation-related keyword trends computed using the same metric, with the top eight keywords (TOP_K_METH = 8) selected based on average frequency (e.g., CpG island, hypermethylation, DNMT, bisulfite sequencing). (**d**) Variability trends quantified using two complementary measures: (i) document-level occurrence rate of predefined variability-related expressions, and (ii) normalized variability score calculated as the frequency of variability-related terms (“variability,” “heterogeneity,” “dynamic,” “plasticity”) divided by total word count per document. Variability-related expressions include chromatin variability, epigenetic plasticity, and chromatin state variation.

This contextual stability suggests that epigenomic research has developed through shared technical conventions and standardized descriptive frameworks. Such linguistic regularity supports reliable feature extraction and facilitates the construction of structured knowledge representations for computational analysis.

Table 4. Integrated concordance and contextual pattern analysis of epigenomic terminology. This table summarizes recurring contextual patterns across genomic loci, experimental methods, chromatin-related expressions, and methylation terms.

Analysis Type	Focus	Description
Genetic Loci Concordance	Chromosomal loci patterns	Concordance contexts showing disease associations and oncogene references near loci expressions (e.g., 18q21.33, 12p13.33, 8q24.21)
Assay Keyword Concordance	Epigenomic experimental methods	Contextual usage of assay-related terms (e.g., 5C, DNase-seq, WGBS, RNA-seq) within methodological descriptions
Loci Before/After Patterns	Lexical proximity around loci	Words surrounding loci expressions indicate genomic positioning and disease relevance
Chromatin Before/After Patterns	Chromatin-related contexts	Surrounding terms highlight genomic annotation, structural regions, and regulatory elements
Assay Before/After Patterns	Experimental method contexts	Assay keywords frequently co-occur with regulatory, profiling, and binding-related terminology
Methylation Before/After Patterns	Epigenetic modification context	Methylation-related terms appear alongside regulatory, silencing, and modification-related expressions

4.5. Temporal Evolution of Epigenomic Research

Temporal trend analysis reveals a significant increase in publication volume beginning around 2010, coinciding with the widespread adoption of high-throughput sequencing technologies, as in Table 5. This period marks a transition toward data-intensive and integrative epigenomic research.

Table 5. Temporal trends of variability-related indicators across five-year intervals. Documents were grouped into five-year bins within the range 1990–2025. Only bins with ≥10 documents were included. Variability-related content was identified using predefined phrase-level patterns (e.g., chromatin variability, epigenetic heterogeneity). Two quantitative metrics were computed: (1) document-level occurrence rate (Variability Document Rate), defined as the proportion of documents containing at least one variability-related expression, and (2) normalized variability score (Mean Normalized Variability Score), calculated as the frequency of variability-related terms normalized by total document length and averaged within each time bin. Publication year was extracted using regular expression matching (19\d{2}|20\d{2}) and assigned based on the most frequent year detected in each document.

Year (5-Year Bin)	Number of Documents	Variability Document Rate	Mean Normalized Variability Score
1990	15	0	0.000413
1995	19	0.105263	0.000402
2000	47	0	0.00057
2005	67	0.044776	0.000686
2010	392	0.07398	0.000714
2015	1525	0.086557	0.000623
2020	3773	0.048237	0.000618

Longitudinal analysis of keyword usage indicates a shift in conceptual emphasis from relatively static regulatory models to frameworks highlighting heterogeneity, plasticity, and dynamic regulation. These trends reflect an increasing focus on context-dependent and time-dependent biological processes.

4.6. Integrated Interpretation of Literature-Level Patterns

Taken together, the results indicate that epigenomic research forms a structured and interconnected knowledge system rather than a collection of isolated studies [31,32]. Multiple analytical dimensions—including epigenetic mechanisms, chromatin variability, gene–epigenetic relationships, disease associations, and temporal dynamics—coexist within a unified conceptual framework. Table 6 summarizes how these components are linked at the literature level and how they collectively support a biomarker-oriented interpretation of epigenomic research.

Table 6. Summary of biological implications and research linkages derived from literature-level text mining. This table integrates key analytical components, including epigenetic mechanisms, chromatin variability, gene–epigenetic relationships, and environmental effects, and describes their roles in the broader conceptual framework of epigenomic research.

Text Mining Component	Core Literature-Level Insight	Link to Chromatin Variability Perspective	Extraction Logic
Epigenetic mechanism	Mechanistic basis of epigenetic regulation	Biological foundation of chromatin variability	Predefined epigenetic mechanism keywords extracted via case-insensitive regular expression-based pattern matching applied to full-text corpus
Chromatin variability expressions	Linguistic operationalization of variability concept	Conceptual validation of chromatin variability	Predefined variability-related expressions identified through case-insensitive keyword-based pattern matching in full text
Gene–epigenetic relationships	Gene-specific regulatory modulation patterns	Functional instantiation of variability	Sentence-level segmentation followed by identification of co-occurrence between gene names and DNA methylation-related expressions
Cell-type specificity	Context-dependent regulatory variation	Contextual interpretation of chromatin variability	Predefined cell-type terminology matched within full-text corpus to identify cell-specific epigenetic contexts
Disease association	Clinical relevance of epigenetic regulation	Translational dimension of variability	Co-occurrence detection of predefined disease terms and epigenetic expressions within full text
Temporal dynamics	Time-dependent regulatory changes	Dynamic nature of chromatin variability	Identification of predefined temporal keywords to detect longitudinal epigenetic contexts
Environmental effects	Non-genetic influences on epigenetic regulation	Environmental modulation of variability	Detection of predefined environmental terms within full-text corpus
Semantic network structure	Conceptual connectivity across literature	System-level interpretation	Construction of co-occurrence networks by linking concept nodes appearing within shared textual contexts
Literature variability score	Quantitative prominence of variability discourse	Independent validation indicator	Frequency aggregation of predefined variability-related keywords across corpus

These components frequently co-occur across studies, forming stable semantic networks that reflect shared research practices and conceptual organization. The presence of consistent variability-related terminology further supports the interpretation of epigenomic regulation as a dynamic and context-dependent system.

4.7. Integrating Literature Patterns with Epigenomic Knowledge

The literature analysis reveals several consistent biomarker-related signals. DNA methylation emerges as the most widely studied epigenomic biomarker across diverse disease domains, including cancer, metabolic disorders, neurological conditions, and aging-related processes. Recurrent experimental methodologies form a stable technical foundation for biomarker discovery, suggesting the presence of standardized analytical pipelines.

Importantly, certain epigenomic features repeatedly appear across multiple disease contexts, indicating their potential as cross-domain biomarkers. These recurring patterns suggest that literature-level frequency, co-occurrence, and contextual stability can serve as criteria for biomarker prioritization.

Chromatin state variability emerges as a key conceptual component, reflecting dynamic regulatory processes shaped by epigenetic plasticity, aging, and disease-related regulatory change [33–35]. Rather than representing static markers, epigenomic states capture regulatory plasticity and system-level adaptation. From this perspective, variability-related features may serve as dynamic indicators for longitudinal health monitoring.

Figure 4 presents a sample segment extracted from the genome-wide chromatin variability tracks, illustrating how variability patterns appear within a representative genomic interval (21,256,927–21,358,256 bp). As shown in Figure 4, chromatin state variability is distributed non-uniformly across human chromosomes, supporting the interpretation that epigenomic plasticity is concentrated in specific genomic regions rather than being evenly distributed across the genome.

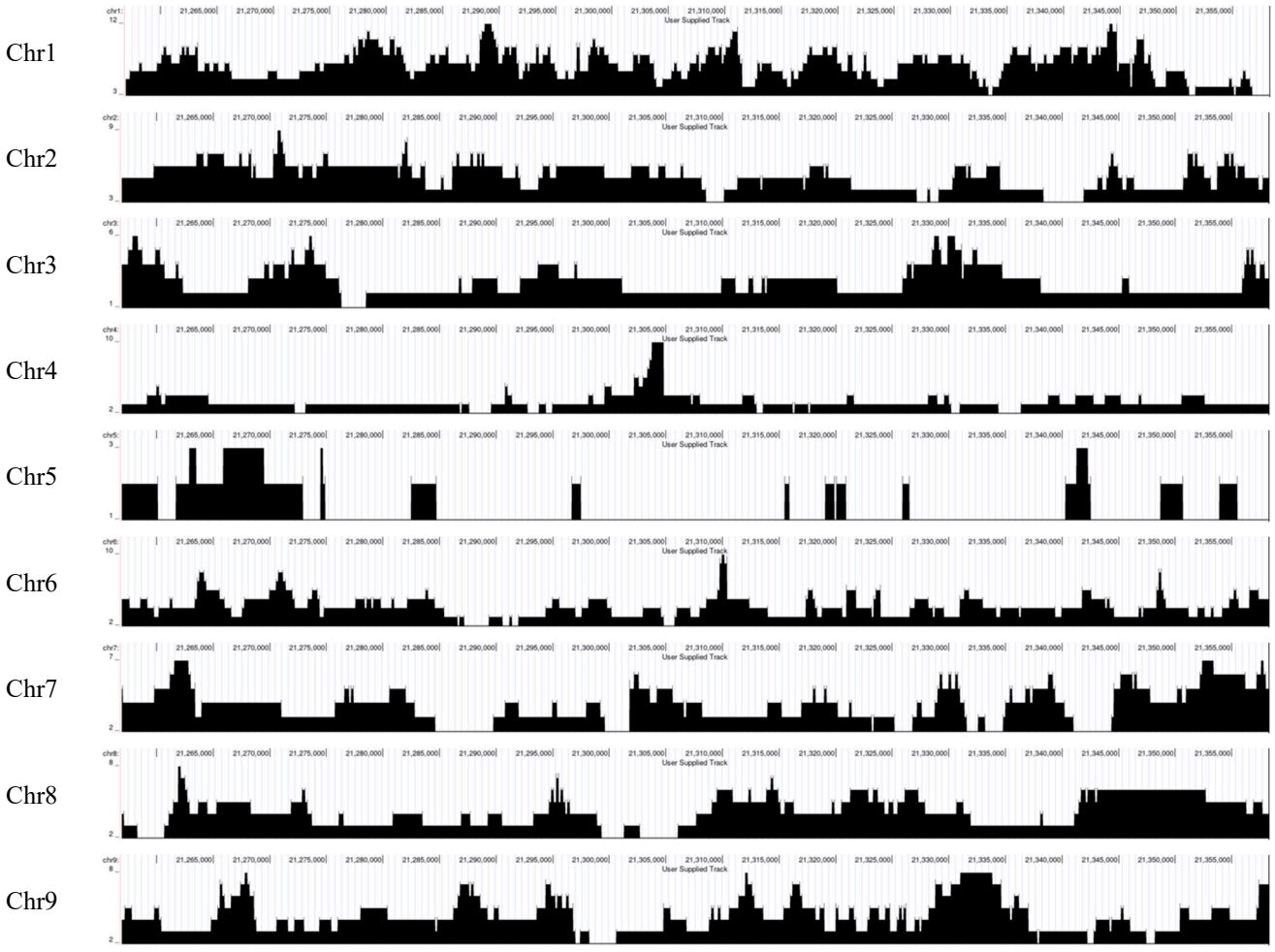


Figure 4. *Cont.*

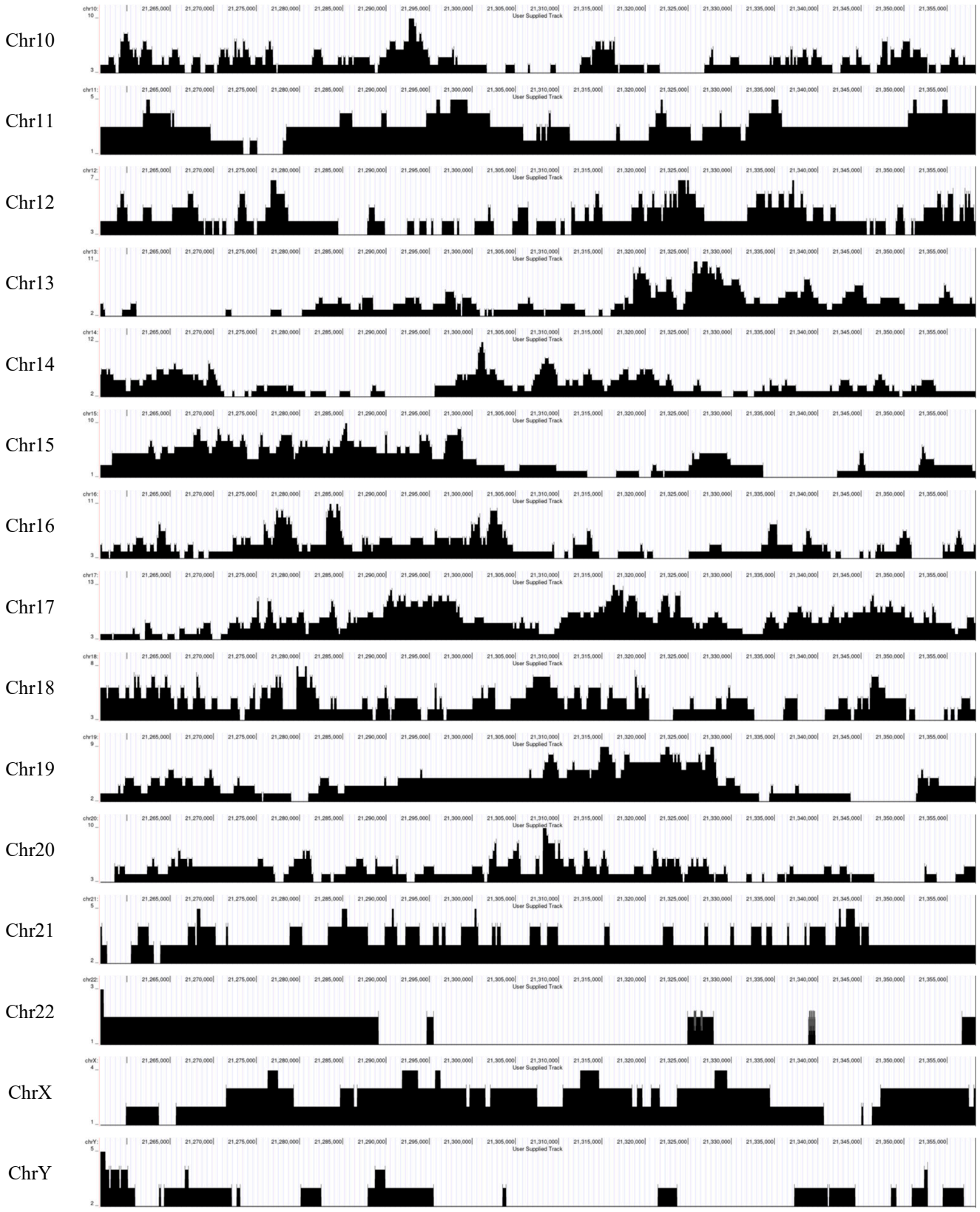


Figure 4. Genome browser view of chromatin state variability across a representative genomic region (21,256,927–21,358,256 bp) . Chromatin variability has been quantified by counting the number of distinct chromatin states within a 200 bp genomic region. The track illustrates variability signals along genomic coordinates, where fluctuations in signal intensity reflect differences in chromatin state assignments across multiple epigenomes. Regions with higher signal variation indicate increased chromatin state diversity, whereas relatively uniform regions suggest stable chromatin configurations.

Although SNP-related terminology frequently co-occurs with chromatin and regulatory elements, such patterns should be interpreted cautiously, as they do not directly imply causal relationships between genetic variation and epigenomic dynamics. Beyond chromatin variability and SNPs, additional chromatin-associated datasets (e.g., histone modifications and chromatin accessibility) can be incorporated, linking multiple epigenomic layers to support biomarker discovery.

Overall, this review suggests that literature-level text mining can serve as a complementary approach to biomarker research by identifying recurrent signals across studies. By leveraging frequency patterns, contextual relationships, and temporal stability, the framework surveyed here supports the systematic prioritization of candidate biomarkers for further investigation and possible future integration into predictive health monitoring systems [31,32]. Taken together, these literature-level patterns do not merely describe the distribution of research topics, but reveal consistent structural properties of epigenomic biomarker knowledge, including recurring associations, contextual stability, and temporal persistence. Such characteristics are directly relevant to how biomarkers are interpreted and utilized, particularly in contexts that require robustness, reproducibility, and longitudinal consistency. In this regard, the observed patterns provide a conceptual basis for considering how literature-derived epigenomic signals may inform the design and interpretation of biomarker-driven health monitoring approaches.

Building on the structured patterns identified in the literature, this section discusses how epigenomic biomarkers—characterized by stability, contextual consistency, and dynamic regulatory features—may be interpret-ed within a broader translational context, particularly in relation to emerging AI-driven health monitoring frame-works.

5. Translational Perspective: AI-Driven Health Monitoring

Recent clinical and translational literature indicates that epigenetic biomarkers are increasingly relevant not only to disease diagnosis but also to the longitudinal assessment of health status. Within the body of work surveyed in this review, DNA methylation signatures have been used to estimate biological aging and health risk, while blood-based methylation assays have demonstrated clinical feasibility for cancer detection, molecular classification, and recurrence-related assessment in several settings [3,4,36–40]. More recent reviews have further emphasized the growing clinical promise of epigenetic biomarkers across cancer, aging, and common complex diseases, while also highlighting the remaining translational gap between biomarker discovery and routine implementation [41–43]. From this review perspective, epigenetic biomarkers may be understood as a plausible molecular foundation for future health monitoring frameworks, particularly when interpreted as relatively stable signals that complement higher-frequency physiological data.

5.1. From Literature Patterns to Health Monitoring Design

Representative clinical studies in epigenetics have already demonstrated that DNA methylation and related regulatory biomarkers can support disease screening, molecular classification, prognostic stratification, and recurrence-oriented assessment in real clinical settings. In aging research, methylation-based predictors have provided quantitative measures of biological age and health risk [3,4]. In oncology, targeted methylation analysis of circulating cell-free DNA has shown feasibility for multi-cancer detection and tissue-of-origin prediction, while methylated SEPT9 has been investigated in colorectal cancer screening, diagnosis, and recurrence monitoring [36–40].

However, most of these studies have been designed for discrete diagnostic, prognostic, or surveillance purposes rather than for continuous health monitoring. Their importance in the present context lies in showing that epigenetic signals can function as clinically

interpretable and repeatedly measurable molecular indicators. When viewed collectively, these case studies suggest that epigenetic biomarkers are not limited to one-time classification tasks, but may also provide stable molecular baselines for repeated assessment over time [41,43].

Accordingly, the contribution of literature-level synthesis is not to claim that an operational health monitoring system has already been validated, but to organize existing clinical and translational evidence into monitoring-relevant biomarker classes. Such an approach may help identify which epigenetic signals are most suitable for repeated sampling, multimodal integration, and longitudinal interpretation, thereby supporting a more systematic transition from diagnosis-centered biomarker studies toward future AI-driven health monitoring frameworks [41,42].

The literature-level synthesis presented in this review provides more than a descriptive overview of epigenomic research. By identifying recurrent biomarker types, stable experimental configurations, and consistent associations between epigenomic mechanisms and disease processes, the analysis establishes a structured knowledge base for translational applications.

As summarized in Figures 1 and 3, this review organizes dispersed findings across the literature into coherent biomarker-relevant patterns that may inform translational interpretation. These structured patterns form a basis for translating epigenomic knowledge into system-level design principles for AI-driven health monitoring.

In this context, epigenomic biomarkers—particularly DNA methylation signatures and chromatin variability indicators—can be interpreted as molecular-level signals that complement physiological measurements derived from digital health technologies [44,45]. The integration of these heterogeneous data types enables the construction of unified monitoring frameworks that combine biological context with real-time physiological dynamics.

Figure 5 conceptually outlines one possible integration pathway linking molecular biomarkers with digital physio-logical monitoring and AI-based inference layers. This representation serves to contextualize how such elements may be combined within emerging research directions, rather than to describe a finalized or experimentally vali-dated system.

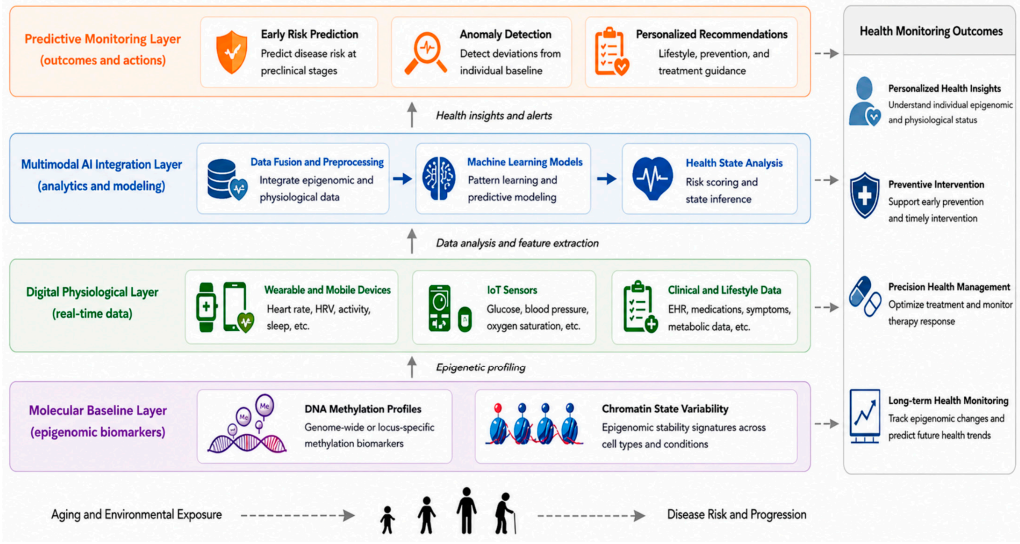


Figure 5. Conceptual architecture of an AI-driven health monitoring system integrating epigenomic biomarkers. This figure presents a multi-layer framework combining molecular biomarkers (e.g., DNA methylation and chromatin variability) with real-time physiological data from wearable devices, integrated through AI models for predictive and personalized health monitoring. This architecture is intended as a conceptual and illustrative example to demonstrate potential system integration, rather than a fully implemented or validated operational framework.

The molecular baseline layer captures DNA methylation and chromatin variability as indicators of regulatory states. These features are combined with real-time signals from wearable devices, IoT sensors, and clinical records in the digital physiological layer. The multimodal AI layer performs data fusion and predictive modeling to assess individual health status. The predictive monitoring layer then generates actionable outputs, including early risk prediction, anomaly detection, and personalized health recommendations, enabling data-driven and individualized healthcare.

5.2. Multi-Layer Architecture for AI-Driven Health Monitoring

A conceptual architecture for AI-driven health monitoring can be organized into four functional layers, each corresponding to a distinct level of biological and computational abstraction.

(1) Molecular biomarker layer

Epigenomic features, including DNA methylation patterns, histone-related regulatory signals, and chromatin variability, provide relatively stable representations of long-term biological states. These markers reflect the cumulative effects of aging, environmental exposure, and disease susceptibility [44,45].

(2) Physiological sensing layer

Wearable devices and IoT-based monitoring systems generate high-frequency physiological data streams, such as heart rate variability, sleep cycles, physical activity, and metabolic indicators. These signals capture short-term and dynamic fluctuations in health status.

(3) Multimodal integration layer

Machine learning models integrate molecular and physiological data to infer latent health states. By combining slow-changing epigenomic signals with rapidly varying physiological measurements, AI models can contextualize transient physiological variations within long-term biological trajectories [44,45].

(4) Predictive monitoring layer

The integrated outputs support downstream applications, including early disease risk prediction, anomaly detection, and personalized health management. These functions are particularly relevant for preventive medicine and continuous health monitoring.

This layered architecture defines a structured pathway for incorporating literature-derived epigenomic knowledge into operational health monitoring systems.

5.3. Multi-Timescale Integration of Epigenomic Signals

A key advantage of incorporating epigenomic biomarkers into health monitoring systems lies in their ability to anchor physiological data within a multi-timescale framework. Unlike real-time physiological signals, epigenomic features evolve gradually and reflect long-term biological processes.

By integrating these complementary temporal scales, AI-driven monitoring systems can simultaneously capture:

- Long-term trajectories, such as aging progression, cumulative environmental exposure, and disease susceptibility;
- Short-term dynamics, including daily physiological variation and acute stress responses.

This multi-timescale integration enables more robust interpretation of physiological signals and reduces the risk of misclassifying transient fluctuations as clinically significant events.

Within this framework, chromatin variability represents a particularly promising signal, as it reflects regulatory plasticity and system-level adaptability. When combined with SNP-associated information and literature-derived biomarker patterns, such variability-related features may contribute to more nuanced representations of biological state transitions.

6. Limitations and Future Directions

As with any literature-based review, the present framework has several methodological and translational limitations that should be considered when interpreting the findings. In this section, we summarize the main constraints of the corpus design, analytical scope, and translational interpretation, and we outline future directions for extending literature-level epigenomic biomarker research toward more robust and clinically meaningful applications.

6.1. Monitoring Limitations of Literature-Based Analysis

Despite the value of large-scale literature mining for knowledge synthesis, several limitations should be considered when interpreting the findings of this review. First, the literature corpus was constructed primarily from open-access articles available through PubMed and PubMed Central. Although this approach enables reproducibility and large-scale full-text analysis, it may introduce selection bias by excluding non-open-access publications and proprietary datasets. As a result, the analyzed corpus may represent only a subset of the broader epigenomic research landscape.

Second, literature-level text mining captures patterns of scientific discourse rather than direct biological measurements. The frequency, co-occurrence, and contextual relationships of keywords reflect research emphasis and reporting practices, which may not fully correspond to experimentally validated biological significance. Therefore, the identified patterns should be interpreted as indicators of collective research trends rather than definitive evidence of biomarker validity.

Third, the analytical framework relies on natural language processing and statistical modeling techniques, including keyword analysis, topic modeling, and concordance-based methods. The results may be sensitive to preprocessing decisions, query design, and model parameters. For example, topic clustering outcomes may vary depending on the number of topics selected or the specific implementation of the modeling algorithm. These factors may influence the stability and reproducibility of the extracted patterns.

Fourth, the present review does not include standardized benchmark-style evaluation metrics commonly used in machine learning-based text mining, such as precision, recall, or topic coherence scores. As a result, the performance of the text mining pipeline is assessed qualitatively rather than through standardized benchmarking measures.

6.2. Limitations in Translational Interpretation

In addition to methodological constraints, this review also recognizes several limitations in the translational interpretation of literature-derived findings.

The proposed linkage between epigenomic biomarkers and AI-driven health monitoring remains conceptual. Although the study outlines a multi-layer integration framework, it does not include implementation-level validation, such as real-world datasets, prototype systems, or predictive performance evaluation. Therefore, the translational relevance of the framework should be interpreted as a theoretical extension rather than a validated application.

Furthermore, co-occurrence patterns observed in the literature—such as associations between SNPs, chromatin states, and disease-related processes—do not imply causal relationships. These patterns may reflect research interest or reporting conventions rather than underlying biological mechanisms. Additional experimental and computational validation is required to establish causal links between genetic variation, epigenomic regulation, and health outcomes.

Finally, the integration of epigenomic biomarkers with real-time physiological data introduces additional challenges, including data heterogeneity, temporal alignment, and

cross-modal integration. These challenges are not explicitly addressed within the scope of the current study.

6.3. Future Directions for Epigenomic Biomarker Research

Future research can address these limitations through several complementary directions.

First, expanding the literature corpus to include non-open-access publications, preprints, and domain-specific databases would improve coverage and reduce potential sampling bias. Integrating multiple biomedical data repositories may provide a more comprehensive representation of epigenomic research.

Second, incorporating quantitative evaluation metrics into the text mining pipeline would enhance methodological rigor and reproducibility. Future studies may evaluate topic coherence, clustering stability, and keyword extraction performance using standardized benchmarks.

Third, integrating literature mining with experimental and multi-omics datasets represents a critical next step. By combining literature-derived patterns with genomic, transcriptomic, and epigenomic datasets, researchers can validate candidate biomarkers and refine their biological interpretation.

Fourth, more advanced natural language processing techniques, including transformer-based models and contextual embeddings, may improve the accuracy of concept extraction and relationship modeling within biomedical texts. These approaches can capture deeper semantic relationships beyond keyword-level analysis.

Fifth, future work should focus on developing and validating prototype systems for AI-driven health monitoring. This includes the integration of epigenomic biomarkers with wearable sensor data, longitudinal cohort studies, and real-world clinical datasets to evaluate predictive performance and practical feasibility.

6.4. Toward Integrated and Validated Health Monitoring Systems

Ultimately, the integration of epigenomic biomarkers into AI-driven health monitoring requires a transition from literature-based hypothesis generation to experimentally validated and clinically deployable systems. This transition involves not only methodological improvements in text mining but also cross-disciplinary collaboration among computational scientists, biologists, and clinicians.

By addressing current limitations and advancing toward validated multimodal frameworks, future research can bridge the gap between large-scale knowledge synthesis and real-world health monitoring applications.

7. Conclusions

This review presented a literature-level synthesis of epigenomic research supported by AI-driven text mining techniques, with a particular focus on DNA methylation, chromatin variability, and biomarker-related patterns. By surveying a large corpus of full-text biomedical articles, the review identified recurring experimental configurations, stable thematic structures, and consistent contextual patterns across the epigenomic research landscape.

The findings suggest that epigenomic research can be understood as a structured and evolving knowledge system characterized by recurring methodological combinations, parallel research domains, and stable technical terminology. Within this literature-level framework, DNA methylation emerges as a central biomarker across diverse disease contexts, while chromatin variability represents an important conceptual element associated with dynamic regulatory processes.

Importantly, this review highlights the value of literature-level text mining as a complementary approach to biomarker research. Rather than focusing on individual experimental

results, the framework presented here supports the identification and organization of recurrent biomarker-relevant signals across studies and may therefore assist the prioritization of candidate biomarkers for further investigation.

From a translational perspective, the integration of epigenomic biomarkers with AI-driven health monitoring frameworks offers a conceptual pathway toward multimodal and multi-timescale health assessment. By combining molecular-level indicators with real-time physiological data, such frameworks may eventually support more predictive and personalized monitoring strategies. However, these applications remain conceptual and require further validation through experimental and clinical studies.

In summary, this review provides a structured framework for surveying epigenomic literature at scale and demonstrates how literature-derived knowledge may inform biomarker prioritization and future health monitoring architectures. Continued integration of computational synthesis, experimental validation, and clinical application will be essential for advancing epigenomic biomarkers toward practical use in precision health systems.

Author Contributions: Conceptualization, H.-S.P. and S.H.S.; methodology, J.-H.O.; software, J.-H.O. and H.-J.N.; validation, H.-S.P. and H.-J.N.; formal analysis, J.-H.O. and H.-J.N.; data curation, J.-H.O.; writing—original draft preparation, J.-H.O.; writing—review and editing, J.-H.O. and H.-S.P.; visualization, J.-H.O. and H.-J.N.; supervision, H.-S.P. and S.H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5A2A03062819).

Institutional Review Board Statement: Ethical review and approval were waived for this study as it did not involve humans or surveys.

Informed Consent Statement: Not applicable.

Data Availability Statement: This study utilized publicly available datasets of epigenomic and DNA methylation-related research articles retrieved from PubMed and PubMed Central (PMC). Only open-access articles with full text were included. Articles were collected using core keywords and related expressions associated with epigenetics and DNA methylation, without restrictions on publication year, disease type, or experimental design. Full-text materials, including titles, abstracts, and complete article bodies, were obtained in plain-text and structured XML formats using PMC identifiers (PMCIDs). The dataset was processed in the BioC format to enable standardized annotation and efficient large-scale text mining. A total of 6152 full-text articles published before 27 January 2024 were included in the final corpus. The curated dataset and implementation codes supporting this study can be found at https://github.com/nemojh7/Epigenomic-Biomarker-literature-mining (accessed on 30 March 2026). Additional scripts and analysis tools used in this study are available from the authors upon reasonable request.

Acknowledgments: During the preparation of this manuscript, the authors used Generative AI (OpenAI, GPT-5.3) for language editing and partial figure generation (Figures 1 and 5).

Conflicts of Interest: The authors declare no conflicts of interest. The funder had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A. Glossary

Table A1. Key terms used in this study.

Term	Description
ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing)	A high-throughput sequencing method used to assess chromatin accessibility by detecting open regions of the genome, enabling the identification of regulatory elements such as enhancers and promoters.
Biomarker	A measurable biological indicator that reflects physiological states, disease processes, or responses to therapeutic interventions, widely used in diagnosis, prognosis, and health monitoring.
ChIP-seq (Chromatin Immunoprecipitation sequencing)	A sequencing-based technique used to analyze protein–DNA interactions, particularly transcription factor binding and histone modifications, across the genome.
Chromatin	A complex of DNA and proteins, primarily histones, that organizes the genome within the nucleus and regulates gene expression through structural and chemical modifications.
Chromatin State Variability	A measure of how chromatin states change across different cell types, tissues, or conditions, reflecting the dynamic regulatory potential of genomic regions.
Concordance Analysis	A text mining method that examines the contextual usage of specific terms or phrases across documents to identify recurring patterns and semantic structures.
Corpus (Literature Corpus)	A large, structured collection of text documents used for computational analysis, typically consisting of scientific articles in literature mining studies.
DNA Methylation	An epigenetic modification involving the addition of a methyl group to cytosine residues, typically at CpG sites, influencing gene expression without altering the DNA sequence.
Epigenetics	The study of heritable and reversible changes in gene expression that occur without alterations in the underlying DNA sequence, often mediated by chemical modifications of DNA and histones.
Epigenomic Biomarker	A biomarker derived from epigenetic features, such as DNA methylation or chromatin state, used to indicate disease status, biological age, or environmental exposure.
Epigenome	The complete set of epigenetic modifications across the genome, including DNA methylation, histone modifications, and chromatin accessibility patterns.
Full-Text Mining	A computational approach that analyzes the entire content of documents, including abstracts and main text, to extract deeper contextual and structural information.
IoT (Internet of Things)	A network of interconnected devices equipped with sensors and communication capabilities, used in health monitoring systems to collect real-time physiological data.
Latent Dirichlet Allocation (LDA)	A probabilistic topic modeling algorithm used to identify latent thematic structures within large text corpora by grouping words into topics.
Machine Learning	A subset of AI that enables systems to learn from data and improve performance on specific tasks without explicit programming, widely applied in biomedical data analysis.

Table A1. *Cont.*

Term	Description
Multi-omics Integration	The combined analysis of multiple types of biological data (e.g., genomics, epigenomics, transcriptomics) to provide a comprehensive understanding of biological systems.
Next-Generation Sequencing (NGS)	High-throughput sequencing technologies that enable rapid and large-scale analysis of DNA and RNA, forming the basis of modern epigenomic studies.
Topic Modeling	A text mining technique used to discover hidden thematic structures in large document collections by identifying clusters of co-occurring words.
Wearable Devices	Electronic devices worn on the body that continuously collect physiological and behavioral data, such as heart rate, activity levels, and sleep patterns.
WGBS (Whole-Genome Bisulfite Sequencing)	A sequencing technique that provides genome-wide, single-base resolution maps of DNA methylation patterns.

References

1. Esteller, M. Epigenetics in cancer. *N. Engl. J. Med.* **2008**, *358*, 1148–1159. [CrossRef]
2. Baylin, S.B.; Jones, P.A. A decade of exploring the cancer epigenome. *Nat. Rev. Cancer* **2011**, *11*, 726–734. [CrossRef]
3. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **2013**, *14*, R115. [CrossRef] [PubMed]
4. Hannum, G.; Guinney, J.; Zhao, L.; Zhang, L.; Hughes, G.; Sadda, S.; Klotzle, B.; Bibikova, M.; Fan, J.-B.; Gao, Y.; et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* **2013**, *49*, 359–367. [CrossRef]
5. Ernst, J.; Kellis, M. ChromHMM: Automating chromatin-state discovery. *Nat. Methods* **2012**, *9*, 215–216. [CrossRef] [PubMed]
6. Roadmap Epigenomics Consortium; Kundaje, A.; Meuleman, W.; Ernst, J.; Bilenky, M.; Yen, A.; Heravi-Moussavi, A.; Kheradpour, P.; Zhang, Z.; Wang, J.; et al. Integrative analysis of 111 reference human epigenomes. *Nature* **2015**, *518*, 317–330. [CrossRef]
7. Lee, K.E.; Park, H.S. Preliminary testing for the Markov property of the fifteen chromatin states of the Broad Histone Track. *Bio-Med. Mater. Eng.* **2015**, *26*, S1917–S1927. [CrossRef]
8. Lent, H.; Lee, K.E.; Park, H.S. Building the frequency profile of the core promoter element patterns in the three ChromHMM promoter states at 200 bp intervals: A statistical perspective. *Genom. Inform.* **2015**, *13*, 152. [CrossRef] [PubMed]
9. Lever, J.; Krzywinski, M.; Altman, N. Points of significance: Model selection and overfitting. *Nat. Methods* **2016**, *13*, 703–704. [CrossRef]
10. Krallinger, M.; Rabal, O.; Akhondi, S.A.; Pérez, M.P.; Santamaría, J.; Rodríguez, G.P.; Tsatsaronis, G.; Intxaurrondo, A.; López, J.A.; Navas-Delgado, I.; et al. Overview of the BioCreative VI chemical-protein interaction track. *Database* **2017**, *2017*, bax019.
11. Sekimizu, T.; Park, H.S.; Tsujii, J.I. Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts. *Genome Inform.* **1998**, *9*, 62–71.
12. Frijters, R.; Heupers, B.; van Beek, P.; Bouwhuis, M.; van Schaik, R.; de Vlieg, J.; Polman, J. Co-occurrence analysis for gene–disease relationships from biomedical literature. *BMC Bioinform.* **2010**, *11*, 91.
13. Zhu, F.; Patumcharoenpol, P.; Zhang, C.; Yang, Y.; Chan, J.; Meechai, A.; Vongsangnak, W.; Shen, B. Biomedical text mining and its applications in cancer research. *J. Biomed. Inform.* **2013**, *46*, 200–211. [CrossRef] [PubMed]
14. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [CrossRef] [PubMed]
15. Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; Poon, H. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc.* **2021**, *3*, 1–23. [CrossRef]
16. Beltagy, I.; Lo, K.; Cohan, A. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 3615–3620.
17. Wei, C.-H.; Allot, A.; Leaman, R.; Lu, Z. PubTator Central: Automated concept annotation for biomedical full text articles. *Nucleic Acids Res.* **2019**, *47*, W587–W593. [CrossRef]
18. Davis, A.P.; Grondin, C.J.; Johnson, R.J.; Sciaky, D.; McMorran, R.; Wiegers, J.; Wiegers, T.C.; Mattingly, C.J. The comparative toxicogenomics database: Update 2021. *Nucleic Acids Res.* **2021**, *49*, D1138–D1143. [CrossRef] [PubMed]

19. Himmelstein, D.S.; Lizee, A.; Hessler, C.; Brueggeman, L.; Chen, S.L.; Hadley, D.; Green, A.; Khankhanian, P.; Baranzini, S.E. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* **2017**, *6*, e26726. [CrossRef]
20. Nicholson, D.N.; Greene, C.S. Constructing knowledge graphs and their applications in biomedical informatics. *Patterns* **2020**, *1*, 100016.
21. Zhao, S.; Su, C.; Lu, Z.; Wang, F. Recent advances in biomedical literature mining. *Brief. Bioinform.* **2021**, *22*, bbaa057. [CrossRef]
22. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
23. Hearst, M.A. Untangling text data mining. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL), College Park, MD, USA, 20–26 June 1999; pp. 3–10.
24. Feldman, R.; Sanger, J. *The Text Mining Handbook*; Cambridge University Press: Cambridge, UK, 2007.
25. Khabsa, M.; Giles, C.L. The number of scholarly documents on the public web. *PLoS ONE* **2014**, *9*, e93949. [CrossRef]
26. West, J.D.; Jacquet, J.; King, M.M.; Correll, S.J.; Bergstrom, C.T. The role of text in scientific article analysis. *J. Informetr.* **2013**, *7*, 487–499.
27. Johnson, D.S.; Mortazavi, A.; Myers, R.M.; Wold, B. Genome-wide mapping of in vivo protein–DNA interactions. *Science* **2007**, *316*, 1497–1502. [CrossRef]
28. Meissner, A.; Gnirke, A.; Bell, G.W.; Ramsahoye, B.; Lander, E.S.; Jaenisch, R. Reduced representation bisulfite sequencing for DNA methylation analysis. *Nucleic Acids Res.* **2005**, *33*, 5868–5877. [CrossRef]
29. Lister, R.; Pelizzola, M.; Dowen, R.H.; Hawkins, R.D.; Hon, G.; Tonti-Filippini, J.; Ecker, J.R. Human DNA methylomes at base resolution. *Nature* **2009**, *462*, 315–322. [CrossRef]
30. Buenrostro, J.D.; Giresi, P.G.; Zaba, L.C.; Chang, H.Y.; Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling. *Nat. Methods* **2013**, *10*, 1213–1218. [CrossRef] [PubMed]
31. Blei, D.M. Probabilistic topic models. *Commun. ACM* **2012**, *55*, 77–84. [CrossRef]
32. Griffiths, T.L.; Steyvers, M. Finding scientific topics. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5228–5235. [CrossRef] [PubMed]
33. Clark, S.J.; Lee, H.J.; Smallwood, S.A.; Kelsey, G.; Reik, W. Single-cell epigenomics: Powerful new methods for understanding gene regulation and cell identity. *Genome Biol.* **2016**, *17*, 72. [CrossRef]
34. Wang, K.; Liu, H.; Hu, Q.; Wang, L.; Liu, J.; Zheng, Z.; Liu, G.H. Epigenetic regulation of aging: Implications for interventions of aging and diseases. *Signal Transduct. Target. Ther.* **2022**, *7*, 374. [CrossRef]
35. Feinberg, A.P. Phenotypic plasticity and the epigenetics of human disease. *Nature* **2007**, *447*, 433–440. [CrossRef] [PubMed]
36. Liu, M.C.; Oxnard, G.R.; Klein, E.A.; Swanton, C.; Seiden, M.V.; CCGA Consortium. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann. Oncol.* **2020**, *31*, 745–759. [CrossRef] [PubMed]
37. Klein, E.A.; Richards, D.; Cohn, A.; Tummala, M.; Lapham, R.; Cosgrove, D.; Chung, G.; Clement, J.; Gao, J.; Hunkapiller, N.; et al. Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Ann. Oncol.* **2021**, *32*, 1167–1177. [CrossRef]
38. Church, T.R.; Wandell, M.; Lofton-Day, C.; Mongin, S.J.; Burger, M.; Payne, S.R.; Castaños-Vélez, E.; Blumenstein, B.A.; Rösch, T.; Osborn, N.; et al. Prospective evaluation of methylated SEPT9 in plasma for detection of asymptomatic colorectal cancer. *Gut* **2014**, *63*, 317–325. [PubMed]
39. Sun, J.; Fei, F.; Zhang, M.; Li, Y.; Zhang, X.; Zhu, S.; Zhang, S. The role of mSEPT9 in screening, diagnosis, and recurrence monitoring of colorectal cancer. *BMC Cancer* **2019**, *19*, 450. [CrossRef]
40. Li, R.; Chen, J.; Shen, X.; Lin, Y.; Tang, J.; Xiong, G.; Zhang, K.; Xiang, M.; Xie, L.; Hu, F. A study of the clinical significance of mSEPT9 in monitoring recurrence and prognosis in patients with surgically treated colorectal cancer. *PLoS ONE* **2024**, *19*, e0312676. [CrossRef]
41. Wisman, G.B.A.; Wojdacz, T.K.; Altucci, L.; Rots, M.G.; DeMeo, D.L.; Snieder, H. Clinical promise and applications of epigenetic biomarkers. *Clin. Epigenet.* **2024**, *16*, 192. [CrossRef]
42. Pharo, H.; Vedeld, H.M.; Sjurgard, I.V.; Pinto, R.; Lind, G.E. From concept to clinic: A roadmap for DNA methylation biomarkers in liquid biopsies. *Oncogene* **2025**, *44*, 4814–4831. [CrossRef]
43. Xu, Z. DNA methylation-based health predictors. *Epigenomics* **2025**, *17*, 1083–1090. [CrossRef]
44. Topol, E.J. High-performance medicine. *Appl. Sci.* **2019**, *25*, 44–56.
45. Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; DePristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; Dean, J. Guide to deep learning in healthcare. *Appl. Sci.* **2019**, *25*, 24–29. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.