Review

Not peer-reviewed version

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

Ji-Hye Oh , Hee-Jo Nam , Su-Hyun Seo , Hyun-Seok Park *

Posted Date: 1 April 2026

doi: 10.20944/preprints202604.0029.v1

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

 doi:10.20944/preprints202604.0029.v1

Review

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

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

[1] Department of Computer Science & Engineering, ELTEC College of Engineering, Ewha Womans University, Seoul 03760, Republic of Korea

[2] Seoul National University Bundang Hospital, 82, Gumi-ro 173beon-gil, Bundang-gu, Seongnam-si, Gyeonggi-do 13620, Rep, of Korea

[3] Department of Computer Science & Engineering, College of AI, Ewha Womans University, Seoul 03760, Korea, Department of Computational Medicine, Ewha Womans University, Seoul 03760, Korea

* Correspondence: neo@ewha.ac.kr (H.-S.P.); Tel.: +82-1049110564 (J.-H.O.); +82-232773513 (H.-S.P.)

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

1. Introduction

1.1. Epigenomic Regulation and Biomarkers in Human Disease

Cancer and many other complex diseases arise through the accumulation of genetic and epigenomic alterations that disrupt normal cellular regulation. While traditional models of tumorigenesis have emphasized genetic mutations such as the activation of oncogenes or the inactivation of tumor suppressor genes, growing evidence demonstrates that epigenomic regulation also plays a fundamental role in disease development and progression. Epigenomic mechanisms regulate gene expression without altering the underlying DNA sequence and include processes such

 doi:10.20944/preprints202604.0029.v1

as DNA methylation and histone modifications (Jones & Baylin, 2002; Bird, 2007). Among these mechanisms, DNA methylation is one of the most extensively studied epigenomic modifications. Aberrant hypermethylation in the promoter regions of tumor suppressor genes is frequently associated with transcriptional silencing and has been widely implicated in tumorigenesis. In addition, histone modifications—including acetylation, methylation, and phosphorylation—alter chromatin structure and influence transcriptional regulation. Importantly, epigenomic alterations are not limited to cancer but are also associated with a wide range of complex diseases, including immune-related and neurological disorders (Bird, 2007; Portela & Esteller, 2010; Feinberg, 2007). These observations highlight the growing importance of epigenomic mechanisms as potential biomarkers for disease detection, prognosis, and therapeutic monitoring.

1.2. Rapid Expansion of Epigenomic Research

With the development of high-throughput sequencing and genome-wide epigenomic profiling technologies, epigenomic research has expanded rapidly across multiple biological systems and disease contexts. As a result, the volume of related scientific literature has increased dramatically over the past two decades. This rapid expansion presents new challenges for researchers attempting to synthesize existing knowledge, identify emerging research themes, and understand long-term methodological and conceptual developments within the field. Although numerous review articles have examined epigenomic mechanisms within specific diseases or biological pathways, relatively few studies have explored the evolution of the epigenomic research landscape itself. A literature-level perspective is therefore increasingly important for identifying recurring conceptual frameworks, stable methodological patterns, and long-term research trajectories across studies.

1.3. AI and Literature-Level Analysis of Biomedical Research

Recent advances in artificial intelligence (AI) and text mining technologies provide powerful tools for analyzing large-scale biomedical literature. Literature-level text mining enables the systematic analysis of large document corpora, allowing researchers to identify recurring concepts, co-occurring research themes, and temporal trends that may not be readily visible in individual studies. By examining the collective structure of scientific publications, such approaches can reveal broader patterns in how epigenomic biomarkers and experimental strategies have been investigated across the research landscape. Rather than evaluating individual experimental findings, literature-driven analysis focuses on identifying structural regularities and conceptual patterns across the body of published research. This perspective can provide valuable insights into how biomarker discovery processes evolve and how methodological practices become established within the scientific community.

1.4. Toward AI-Driven Health Monitoring Using Epigenomic Biomarkers

In recent years, epigenomic biomarkers have attracted increasing attention as dynamic indicators of physiological states, disease progression, and environmental exposure. Unlike static genetic mutations, epigenetic modifications such as DNA methylation can change over time in response to biological and environmental factors, making them particularly suitable for longitudinal health monitoring. The rapid development of AI-driven digital health ecosystems—including wearable monitoring systems, IoT-enabled biosensors, and large-scale biomedical data integration platforms—has further increased the demand for structured biomarker knowledge that can be computationally integrated into health monitoring systems. In this context, systematic synthesis of epigenomic biomarker knowledge derived from large-scale literature analysis may provide a foundational layer for AI-enabled health monitoring architectures. Accordingly, this review applies AI-based text mining techniques to analyze a large corpus of biomedical literature related to epigenomic regulation and DNA methylation. By reconstructing the structural patterns and research trajectories of epigenomic studies at the literature level, we aim to identify biomarker-relevant

 doi:10.20944/preprints202604.0029.v1

insights and discuss how such knowledge can contribute to the development of AI-driven health monitoring frameworks within precision health environments.

2. Epigenomic Biomarkers and Chromatin Variability in Human Health

2.1. Biological and Clinical Significance of DNA Methylation

DNA methylation represents 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(Esteller, 2008). Unlike genetic mutations, epigenomic modifications are reversible and dynamically regulated, enabling them to reflect environmental exposures, aging processes, lifestyle factors, and disease progression. These properties make DNA methylation particularly suitable for biomarker discovery and longitudinal health monitoring. In cancer research, aberrant DNA methylation patterns are closely associated with tumor initiation, progression, and metastasis (Baylin & Jones, 2011). Hypermethylation of tumor suppressor gene promoters is frequently linked to transcriptional silencing and has been widely proposed as a major mechanism of tumorigenesis. DNA methylation biomarkers have therefore been actively investigated for early tumor detection and disease classification, including through non-invasive liquid biopsy approaches. Beyond oncology, methylation alterations have also been associated with metabolic disorders, cardiovascular diseases, and neurological conditions, highlighting the broad applicability of epigenomic biomarkers across multiple disease domains. Advances in epigenomic profiling technologies, including high-throughput sequencing and single-cell epigenomic analysis, have significantly expanded the resolution at which methylation patterns can be characterized. These technological developments have accelerated biomarker discovery and facilitated the identification of disease-associated methylation signatures across diverse tissues and biological contexts.

2.2. DNA Methylation Biomarkers in Aging and Health Monitoring

In addition to disease-specific biomarkers, DNA methylation has emerged as a key molecular indicator of biological aging. Age-associated methylation changes have been widely documented, and several epigenetic clock models have been developed to estimate biological age based on methylation signatures (Horvath, 2013). These models capture not only chronological aging but also the cumulative effects of environmental exposures, lifestyle factors, and disease **processes (**Hannum et al., 2013**).**

From a preventive medicine perspective, epigenetic clocks provide a molecular framework for early risk stratification. Because methylation changes can occur prior to overt clinical symptoms, biological age estimation may enable proactive identification of individuals at elevated risk for age-related diseases. Consequently, DNA methylation biomarkers have been increasingly explored as quantitative indicators for long-term health monitoring, disease risk prediction, and population-level health assessment. These characteristics position epigenomic biomarkers as promising components of next-generation digital health systems. When integrated with clinical data streams and physiological monitoring platforms, dynamic epigenomic indicators may contribute to improved predictive modeling and personalized health management.

2.3. Translational Perspectives and the Need for Literature-Level Synthesis

As epigenomic research continues to expand, there has been a substantial increase in the diversity of experimental methodologies, analytical frameworks, and biological contexts in which DNA methylation is studied. This growing complexity presents challenges for synthesizing biomarker-related insights across individual studies **(**Ioannidis, 2016**).** Although many narrative reviews have examined epigenomic mechanisms within specific diseases, relatively few studies have systematically explored how biomarker-oriented epigenomic research has evolved across the broader scientific literature. A literature-level perspective is therefore increasingly necessary to identify

 doi:10.20944/preprints202604.0029.v1

recurring methodological patterns, stable conceptual frameworks, and long-term research trajectories within this rapidly developing field **(**Smalheiser et al., 2021**).** Such structured synthesis may provide valuable guidance for biomarker prioritization and support the integration of epigenomic knowledge into emerging AI-driven health monitoring frameworks.

2.4. Modeling Chromatin Variability from DNA Sequences

Understanding the determinants of epigenomic variability represents an important step toward identifying functional regulatory elements and potential biomarkers. Recent large-scale epigenomic studies have investigated how DNA sequence features constrain chromatin state variability across cell types and tissues **(**Kundaje et al., 2015**).** Using ChromHMM annotations derived from 127 reference epigenomes, genomic regions can be segmented into short sequence bins and analyzed to quantify the diversity of chromatin states observed across multiple biological contexts(Lee, K. E., & Park, H. S., 2015; Lent, H., Lee, K. E., & Park, H. S., 2015**;** Ernst & Kellis, 2012**).** A key metric introduced in such analyses is the chromatin state variability count, defined as the number of distinct chromatin states detected at a given genomic locus across different epigenomes. This measure enables the classification of genomic regions into relatively stable regulatory elements and highly dynamic regulatory hotspots. Genome-wide analyses reveal that certain loci exhibit frequent chromatin state transitions across cell types, reflecting epigenetically dynamic regulatory regions, whereas neighboring genomic segments often maintain stable chromatin configurations(Lee, K. E., & Park, H. S., 2015). These spatial patterns suggest that regulatory elements are enriched in epigenomic variability, while structurally constrained genomic regions tend to maintain conserved epigenetic states. Recent machine learning approaches further demonstrate that chromatin variability can be partially predicted from DNA sequence features alone. Convolutional neural network models trained on genomic sequences have shown the ability to predict chromatin state variability with meaningful accuracy, indicating that sequence-encoded regulatory information contributes to epigenomic plasticity(Lent, H., Lee, K. E., & Park, H. S., 2015). Together, these findings provide a conceptual bridge between genomic sequence architecture and epigenomic regulation. Modeling chromatin variability therefore offers a scalable framework for identifying regulatory elements and prioritizing candidate epigenomic biomarkers, particularly within noncoding genomic regions implicated in disease.

 doi:10.20944/preprints202604.0029.v1

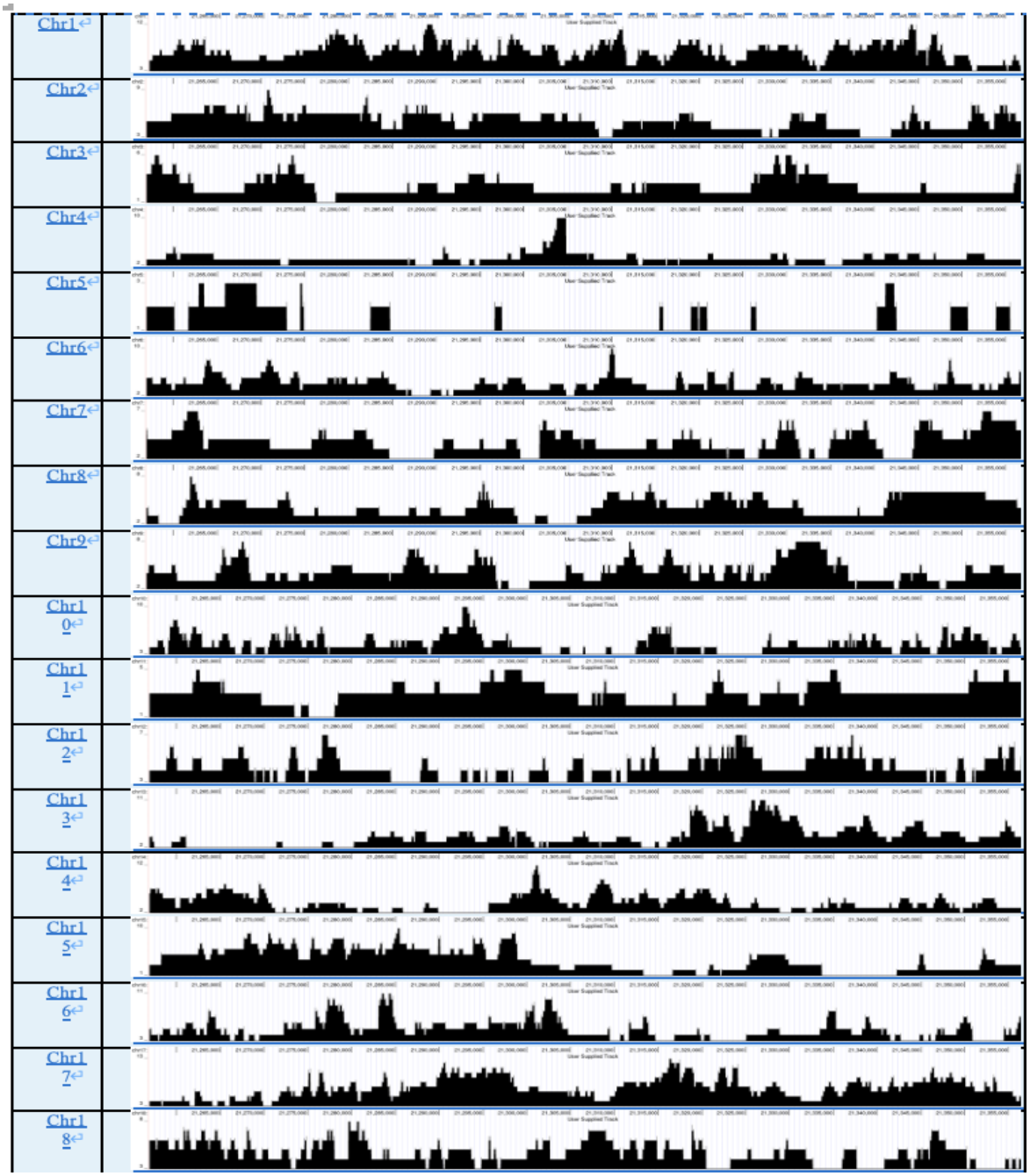


Figure 1. Genome-wide distribution of chromatin state variability across human chromosomes. This figure illustrates the variability of chromatin states across genomic regions, highlighting the distinction between stable regulatory elements and dynamically changing loci across multiple epigenomic contexts.

While chromatin variability reflects dynamic epigenomic regulation across biological contexts, the extent to which underlying genetic variation—particularly disease-associated single nucleotide polymorphisms (SNPs)—contributes to or constrains such variability remains incompletely understood.

Although emerging studies suggest that DNA sequence features may influence chromatin state organization, the direct relationship between disease-associated SNPs and chromatin variability has not yet been systematically established.

 doi:10.20944/preprints202604.0029.v1

3. AI and Text Mining in Biomedical Literature

3.1. AI Applications in Epigenomic Biomarker Research

The rapid expansion of biomedical data has accelerated the adoption of artificial intelligence (AI) and machine learning methods for biomarker discovery and health prediction **(**Zhao et al., 2021**).** In epigenomic research, AI models have been widely applied to DNA methylation data to improve disease classification, predict biological age, and detect early disease signals. In cancer epigenomics, deep learning approaches have enabled the identification of tumor-specific methylation signatures across multiple cancer types, supporting the development of multi-cancer early detection systems. Similarly, AI-based epigenetic clock models have demonstrated improved accuracy in estimating biological age and assessing disease risk compared with traditional statistical approaches. Recent studies have also integrated DNA methylation profiles with inflammatory markers to construct composite health indicators capable of distinguishing healthy individuals from patients with chronic diseases. These developments illustrate the growing role of AI in transforming epigenomic data into actionable health monitoring tools.

3.2. Limitations of Current Biomarker Studies

Despite these advances, most epigenomic biomarker studies have primarily focused on either experimental identification of specific methylation markers or the development of predictive models using molecular datasets. Relatively few investigations have examined the broader structure of epigenomic research itself from a literature-level perspective **(**Zhu et al., 2013**).** In particular, the relationships among disease-specific research domains and their potential implications for biomarker discovery and health monitoring remain insufficiently explored. Understanding these structural patterns requires systematic analysis of the scientific literature rather than isolated examination of individual experimental studies.

3.3. Evolution of Biomedical Literature Mining

Biomedical literature mining has developed through two partially distinct research traditions: biomedical informatics (BMI) and computer science (CS). Comparative analyses suggest that BMI research tends to emphasize interpretability and domain knowledge integration, whereas CS research often focuses on predictive performance and algorithmic generalization **(**Blei et al., 2003**).** Recent studies have emphasized the importance of bridging these perspectives through integrative frameworks that combine domain knowledge with scalable computational methods. Advances in distributed computing and machine learning have further enabled large-scale literature analysis. For example, scalable frameworks such as SparkText have demonstrated the ability to process thousands of biomedical articles in near real time using distributed architectures and automated classification techniques. These technological developments have made high-throughput literature mining increasingly feasible for large biomedical corpora.

3.4. Full-Text Mining for Knowledge Discovery

Traditional narrative reviews rely on selective manual reading, which limits scalability and reproducibility. In contrast, text mining enables systematic analysis of large document collections and facilitates the identification of recurring concepts, relationships, and temporal trends across studies **(**Zhao et al., 2021**).** Recent research has shown that full-text analysis provides substantially richer information than abstract-based mining. Full articles contain detailed methodological descriptions, experimental contexts, and biological relationships that are often absent from abstracts. For example, full-text mining has been used to construct detailed biological interaction networks and curated datasets of protein–protein and genetic interactions derived from PubMed Central articles. These findings highlight the importance of full-text corpora for accurate knowledge extraction in biomedical research.

 doi:10.20944/preprints202604.0029.v1

3.5. Literature-Level Insights for Epigenomic Biomarker Discovery

In the context of epigenomic research, AI-driven text mining enables the systematic identification of frequently studied genomic loci, chromatin regions, experimental methodologies, and methylation-related concepts across diverse disease domains **(**Zhu et al., 2013; Blei et al., 2003**).** By analyzing large-scale literature corpora, such approaches reveal higher-level organizational patterns that cannot be observed through individual studies alone. Literature-level analysis therefore provides a complementary perspective to experimental biomarker research. By organizing dispersed scientific findings into structured knowledge frameworks, AI-assisted literature mining can support biomarker prioritization and facilitate the integration of epigenomic knowledge into emerging digital health systems. Within rapidly expanding research fields such as epigenomics, where methodological diversity and conceptual complexity continue to increase, literature-level synthesis represents an important strategy for transforming fragmented scientific evidence into coherent insights relevant to health monitoring and precision medicine.

4. Literature Mining Framework for Epigenomic Biomarker Discovery

4.1. Data Sources and Corpus Construction

The literature corpus analyzed in this study was constructed from epigenomic and DNA methylation–related research articles retrieved from PubMed and PubMed Central (PMC). To enable large-scale literature analysis, only open-access articles with available full text were included**(**Hearst, 1999; Feldman & Sanger, 2007**).** 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 in order to capture a broad representation of epigenomic biomarker research. Full-text materials—including titles, abstracts, and complete article bodies—were collected using PubMed Central identifiers (PMCIDs). Articles were obtained in both plain-text and structured XML formats to facilitate automated text processing. For large-scale text mining, the study utilized the BioC data format, which provides a standardized XML/JSON representation of biomedical articles and supports efficient computational processing. BioC-based PMC collections are widely used in biomedical text mining because they enable consistent annotation, interoperability, and scalable corpus management. Using this retrieval strategy, a total of 6,152 full-text articles published before January 27, 2024 were included in the final corpus. The resulting dataset provides a structured foundation for literature-level analysis of epigenomic biomarker research.

4.2. Query Pattern Design for Epigenomic Biomarker Retrieval

To construct a comprehensive literature corpus, structured query patterns were designed to capture a broad range of epigenomic biomarker studies. The retrieval strategy combined core epigenetic concepts—such as **DNA methylation, histone modification, and chromatin regulation**—with biomarker-related terminology**(**Zhao et al., 2021**).**

The query design followed a two-level strategy:

- **Baseline queries** were constructed to achieve high recall by broadly capturing literature related to epigenetic biomarkers across diverse disease contexts.
- **Extended queries** incorporated additional constraints, such as disease categories, mechanistic descriptors (e.g., hypermethylation or hypomethylation), clinical cohort indicators, and machine-learning–related terminology.

This two-tier approach balances **corpus coverage and thematic specificity**, allowing the framework to retrieve a comprehensive yet structured collection of biomarker-related studies suitable for large-scale text mining. Table 1 presents structured query patterns designed for broad retrieval of epigenetic biomarker–related literature in large-scale text mining. These baseline queries emphasize high recall by incorporating core epigenetic terms (e.g., DNA methylation, histone

 doi:10.20944/preprints202604.0029.v1

modification, chromatin accessibility) in combination with biomarker-related expressions. The purpose of this design is to ensure comprehensive corpus construction while capturing diverse biomarker research contexts, including diagnostic, prognostic, regulatory, and multi-omics studies.

Table 1. 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

 doi:10.20944/preprints202604.0029.v1

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 (e.g., hypermethylation, hypomethylation), clinical cohort indicators, machine learning–related terminology, and validation-focused terms. By integrating these targeted filters, the retrieval framework improves specificity and prioritizes translationally relevant, mechanistically characterized, and reproducible biomarker studies.

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

 doi:10.20944/preprints202604.0029.v1

Validation-focused studies	(“epigenetic biomarker”) AND validation	Retrieves studies emphasizing reproducibility and independent validation

Together, the two-tier query design balances recall and precision, enabling scalable yet structured corpus generation suitable for literature-level epigenomic biomarker analysis.

4.3. Text Mining and Analytical Pipeline

To analyze structural patterns within the epigenomic literature, multiple complementary text mining techniques were applied to the corpus.

First, keyword frequency analysis and co-occurrence analysis were used to identify frequently mentioned biological concepts, genomic elements, and experimental techniques in epigenomic biomarker research. These methods reveal common research focal points and conceptual associations across the literature.

Second, topic clustering analysis was performed to identify higher-level thematic structures within the corpus. Topic modeling techniques enable the detection of latent research themes and allow the literature to be organized into major thematic domains**(**Blei et al., 2003**).**

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

Finally, temporal trend analysis was conducted to investigate how research themes and methodological approaches have evolved over time within the epigenomic literature.

Together, these analytical components form an integrated literature-level knowledge synthesis framework. By organizing dispersed biomarker evidence into structured thematic and temporal patterns, this framework provides interpretable insights that may support biomarker prioritization and inform downstream AI-based health monitoring systems.

Figure 2 illustrates the overall text mining framework for literature-level analysis of epigenomic and DNA methylation research, highlighting the data acquisition pipeline, analytical modules, and their conceptual linkage to AI-driven health monitoring integration.

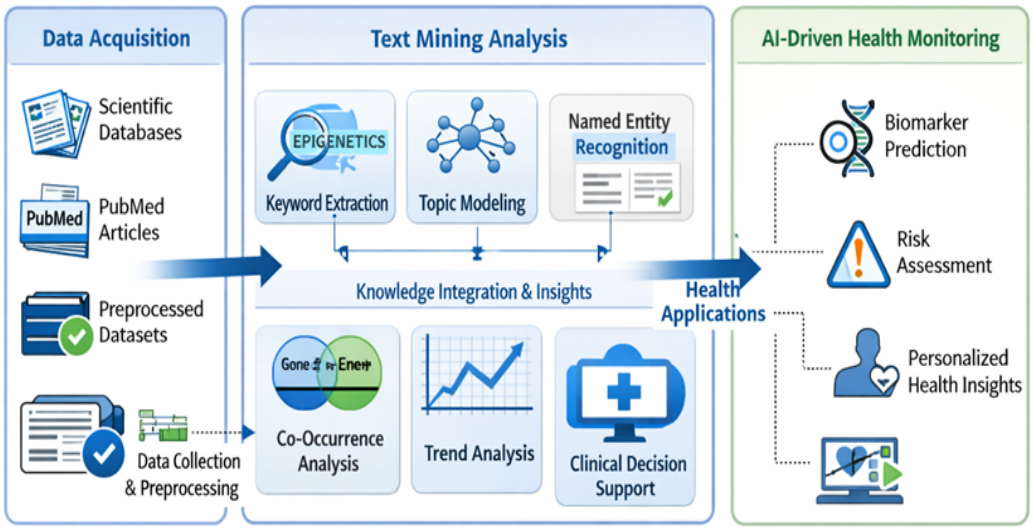


Figure 2. Overview of the literature-level text mining framework for epigenomic biomarker analysis. This figure depicts the overall analytical pipeline, including data acquisition from PubMed and PubMed Central, corpus construction, text mining modules (keyword analysis, topic modeling, concordance analysis, and temporal analysis), and their integration into AI-driven health monitoring applications.

 doi:10.20944/preprints202604.0029.v1

5. Characteristics of the Epigenomic Literature Corpus

5.1. Corpus Scope and Research Coverage

The literature corpus analyzed in this study consists of epigenomic and DNA methylation–related research articles accumulated over several decades. To capture the broad landscape of epigenomic research, the corpus was constructed without restrictions on specific diseases, experimental models, or study designs(Khabsa & Giles, 2014). This inclusive strategy allows the dataset to represent diverse thematic areas and research contexts within the field of epigenomics. As a result, the corpus provides a suitable foundation for examining long-term research trends and structural characteristics across the epigenomic biomarker literature.

5.2. Textual Composition and Data Preprocessing

Each article in the corpus includes complete textual components, comprising the title, abstract, and full-text body(West et al., 2013). The inclusion of full-text content enables more comprehensive analysis compared with abstract-only datasets, as detailed methodological descriptions and contextual information are typically contained within the main article text. To enable cross-document comparison and large-scale text mining, preprocessing procedures were applied to both textual content and associated metadata. These preprocessing steps were designed to reduce formatting heterogeneity across publications and to ensure consistent analytical criteria for subsequent computational analysis.

5.3. Corpus Scale and Analytical Implications

The resulting dataset represents a large-scale biomedical literature corpus containing thousands of full-text articles and tens of millions of words. The scale and structural diversity of the dataset highlight the limitations of traditional manual review approaches and underscore the need for computationally assisted knowledge synthesis. From an analytical perspective, the size of the corpus enables the detection of stable conceptual patterns, thematic clusters, and long-term research trends within the epigenomic literature. Such large-scale analysis provides a systematic foundation for identifying recurring biomarker-related concepts and methodological trajectories relevant to health monitoring research.

6. Major Trends Identified Through Literature-Level Analysis

6.1. Experimental Techniques and Research Design Patterns

Keyword analysis of experimental methodologies revealed that epigenomic research extensively employs high-throughput sequencing and profiling techniques, including RNA-seq, ChIP-seq, DNase-seq, ATAC-seq, as well as DNA methylation profiling approaches such as RRBS and WGBS (Johnson et al., 2007; Meissner et al., 2005; Lister et al., 2009; Buenrostro et al., 2013). These technologies have become representative epigenomic analytical methods, enabling precise measurement of chromatin accessibility, DNA–protein interactions, transcriptomic alterations, and single-base resolution methylation patterns. Rather than being applied independently, these techniques frequently appeared in recurring combinations depending on specific analytical objectives and biological contexts. At the literature level, such recurrent methodological pairings suggest the emergence of relatively standardized research design patterns within the epigenomic field. For example, RNA-seq frequently co-occurred with ChIP-seq analyses aimed at exploring chromatin states and regulatory mechanisms. Similarly, DNase-seq for chromatin accessibility and quantitative DNA methylation profiling techniques were repeatedly observed together within particular experimental contexts. These methodological combinations were selected in relatively consistent ways depending on genomic regions of interest, regulatory mechanisms under investigation, or biological phenomena being studied. This pattern reflects shared analytical

 doi:10.20944/preprints202604.0029.v1

strategies and methodological regularities across the broader epigenomic research landscape. Tables 4 through 3 summarize the usage frequency and combinatorial patterns of these major experimental techniques, providing a structured representation of research design regularities observed across the literature corpus. From an applied AI perspective, the recurrence of specific methodological pairings indicates the presence of stable experimental configurations that could inform automated biomarker selection pipelines and machine learning–based feature integration frameworks. Identifying such standardized design patterns at the literature level may contribute to the development of reproducible and interpretable health monitoring architectures that integrate multi-omics evidence streams.

Table 3. Summary statistics of the epigenomic and DNA methylation literature corpus. This table provides an overview of the corpus used for analysis, including the total number of articles, total word count, and average number of words per article, highlighting the large-scale nature of the dataset.

Total No. of words	62,360,028
No. of articles retrieved	6,152
No. of words per article	10,136.5

Table 4. Journal- and document-level distribution of epigenomic experimental method keywords. This table summarizes the frequency and co-occurrence of experimental method keywords across journals and individual documents, illustrating common methodological patterns and their distribution within the literature corpus.

Level	Description	Representative Examples
Journal Level (Frequency)	Journals in which epigenomic experimental method keywords most frequently appear	The Journal of Biological Chemistry (111); Physical Review B (102); The Cochrane Database of Systematic Reviews (88)
Journal–Keyword Association	Frequently co-occurring keywords within selected journals	Science (predictions, Computationall); Gene (5C); PNAS (predictions)
Document Level	Documents containing multiple epigenomic experimental method keywords	Epi_file-10000085.txt (5C, Hi-C, DNase-seq, ATAC-seq, ChIP-seq, RNA-seq); Epi_file-10000364.txt (DNase-seq, ATAC-seq, ChIP-seq, RNA-seq)

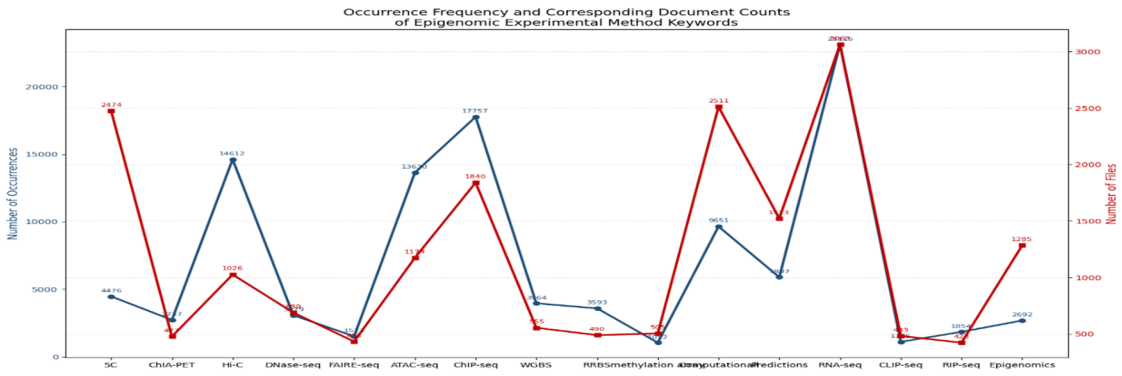


Figure 3. Frequency and document coverage of epigenomic experimental method keywords. The blue line represents the total frequency of each keyword across the corpus, while the red line indicates the number of

 doi:10.20944/preprints202604.0029.v1

documents in which each keyword appears, demonstrating the prevalence and distribution of experimental techniques.

6.2. Topic Structure of Epigenomic Research

Topic clustering analysis revealed that the epigenomic literature does not converge into a single dominant research trajectory but rather consists of multiple parallel research domains. This finding indicates that epigenomic research has evolved in a multilayered manner across diverse biological contexts and investigative objectives. The major topic clusters identified at the literature level included (1) disease-centered research, (2) chromatin structure and transcriptional regulation studies, (3) integrative analyses combining epigenomic and transcriptomic data, and (4) cancer-related epigenomic research. These clusters share common conceptual frameworks and analytical techniques while forming distinct focal areas within a parallel structural configuration. Overall, this topic structure suggests that epigenomic research has expanded through the parallel development of multiple interconnected domains, accumulating knowledge in a complementary rather than linear manner. The identification of parallel yet interconnected research domains provides a structured conceptual map that may serve as an organizational scaffold for AI-driven knowledge graphs and health monitoring models. By aligning biomarker evidence within these thematic clusters, machine learning systems can potentially incorporate domain-specific contextual weighting into predictive health analytics.

Table 5. Topic cluster composition and representative keywords identified through topic modeling. This table presents the major topic clusters identified from the corpus, including the number of documents per cluster and representative keywords, reflecting the thematic structure of epigenomic research.

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, DNA, CpG, methylated, epigenetic, site, ages
Cluster 3	1119	mice, RNA, mRNA, p, supplementary, h, mM, N
Cluster 4	824	Cancer, tumor, patients, immune, Breast, survival, Breast, p
Cluster 5	901	Chromatin, enhancer, peak, genome, transcription, DNA, binding, accessibility
Cluster 6	733	RNA, Cancer, DNA, patients, disease

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 April 2026 doi:10.20944/preprints202604.0029.v1

Figure 4. Heatmap of associations between epigenomic experimental methods and methylation-related keywords. This figure visualizes the co-occurrence patterns between experimental techniques and methylation-related terms, revealing structured relationships among commonly used analytical approaches.

6.3. Contextual Stability of Epigenomic Terminology

Concordance-based analysis revealed that genomic loci, chromatin coordinates, experimental methodologies, and DNA methylation-related concepts were consistently embedded within relatively stable narrative structures across the literature corpus. Core technical terms did not appear randomly but were systematically employed within recurring descriptive frameworks. Such contextual consistency suggests that epigenomic research has accumulated through relatively predictable technical conventions and discourse patterns. Core concepts and experimental methodologies were repeatedly articulated within similar descriptive structures across diverse biological contexts. These findings indicate that the epigenomic literature represents not merely an aggregation of individual study outcomes, but a structured knowledge system characterized by shared technical vocabulary and discursive conventions. This linguistic stability is particularly advantageous for AI-driven health monitoring applications, as stable terminology facilitates reliable feature extraction, ontology construction, and automated knowledge integration within digital health platforms.

Table 6. Integrated concordance and contextual pattern analysis of epigenomic terminology. This table summarizes concordance-based analyses of genomic loci, experimental methods, chromatin-related expressions, and methylation terms, highlighting recurring contextual usage patterns across the literature.

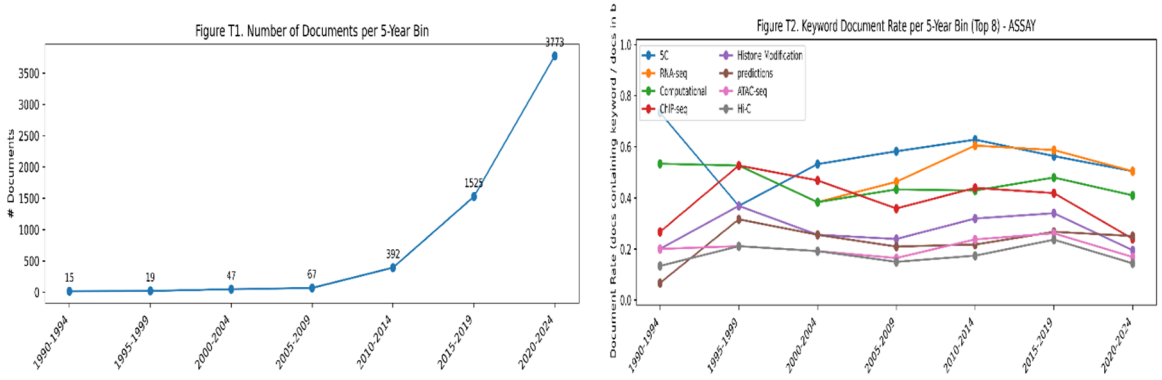
Analysis Type	Focus	Description
Genetic Loci Concordance	Chromosomal loci patterns	Concordance contexts showing disease associations and oncogene references near loci expressions (e.g., 18q21.33, 12p13.33, 8q24.21)
Assay Keyword Concordance	Epigenomic experimental methods	Contextual usage of assay-related terms (e.g., 5C, DNase-seq, WGBS, RNA-seq) within methodological descriptions

 doi:10.20944/preprints202604.0029.v1

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

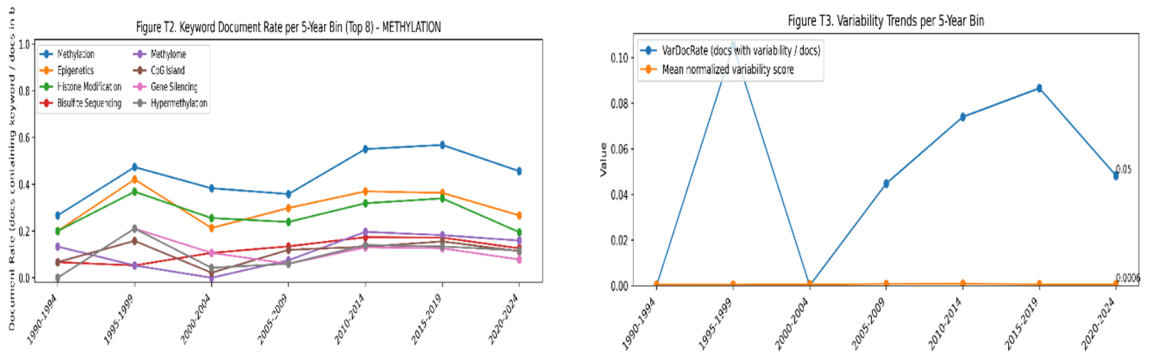
6.4. Temporal Evolution of Epigenomic and DNA Methylation Research

Temporal trend analysis revealed a marked increase in publication volume beginning around 2010. This inflection point coincides with the widespread adoption of high-throughput sequencing technologies and the emergence of epigenomic data as a central analytical resource in biomarker discovery research. Figures 5(b) through 5(d) illustrate longitudinal changes in publication counts and keyword usage frequencies across five-year intervals. The sustained growth in publication output indicates that epigenomic research has expanded across diverse biological and medical domains. Analysis of variability-related terminology revealed shifting conceptual emphases over time. Expressions such as “heterogeneity” and “plasticity” exhibited differential prominence across distinct intervals. This pattern suggests a gradual transition from static regulatory models toward frameworks emphasizing intercellular diversity, state-dependent variation, and dynamic regulation. Tables 7 quantitatively demonstrate changes in research focus and conceptual orientation within epigenomic and DNA methylation studies. From a health monitoring perspective, this temporal shift toward variability-oriented concepts aligns with the increasing demand for dynamic, longitudinal risk prediction models. The emphasis on heterogeneity and adaptive regulation parallels the conceptual foundations of AI-based personalized monitoring systems that account for inter-individual variability and temporal state transitions.



(**a**) (**b**)

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 April 2026 doi:10.20944/preprints202604.0029.v1



(c) (d)

Figure 5. Temporal trends in epigenomics and DNA methylation research (1990–2024). (a) Number of publications over time; (b) trends in assay-related keywords; (c) trends in methylation-related keywords; and (d) trends in variability-related expressions, illustrating the rapid expansion and conceptual evolution of the field.

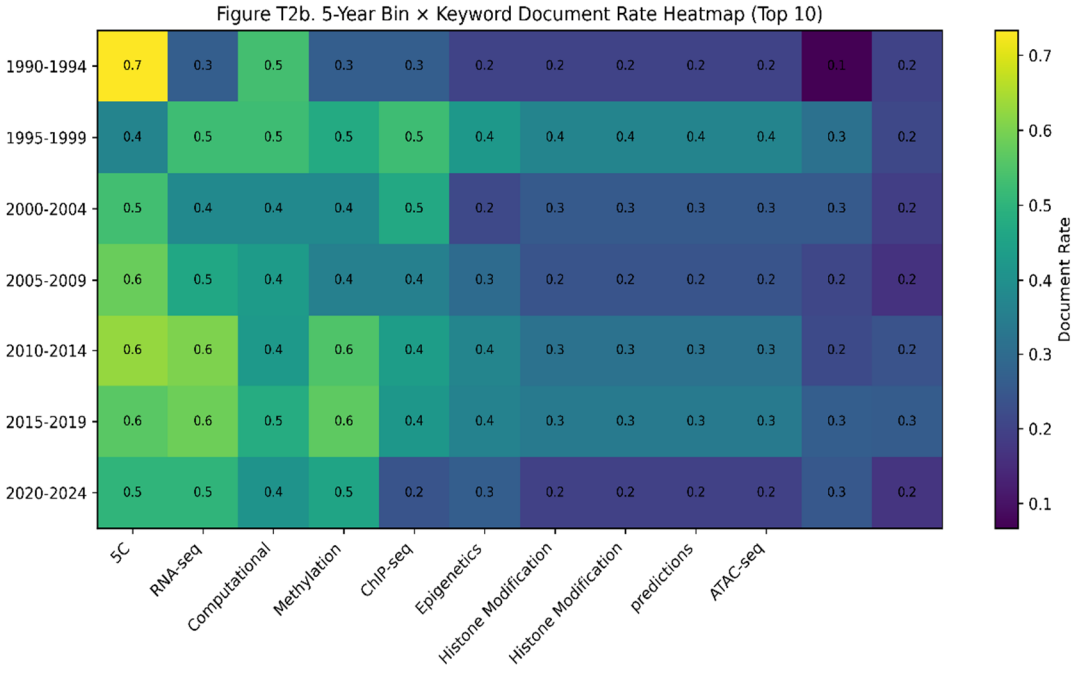


Figure 6. Heatmap of keyword document rates across five-year intervals. This figure shows the relative frequency of top-ranked epigenomic and methylation-related keywords over time, highlighting shifts in research focus and thematic evolution.

Table 7. Temporal trends of variability-related indicators across five-year intervals. This table presents quantitative changes in variability-related concepts over time, including document counts and normalized variability scores, reflecting the evolving emphasis on epigenomic dynamics.

year_bin	n_docs	var_doc_rate	var_score_norm_mean
1990	15	0	0.000413
1995	19	0.105263	0.000402
2000	47	0	0.00057
2005	67	0.044776	0.000686

 doi:10.20944/preprints202604.0029.v1

2010	392	0.07398	0.000714
2015	1525	0.086557	0.000623
2020	3773	0.048237	0.000618

7. Integrated Interpretation of Literature-Level Text Mining Results

7.1. Integrated Framework of Epigenomic Literature Patterns

This section synthesizes the results of the literature-level text mining analyses and presents an integrated interpretation of the structural patterns identified across the epigenomic research corpus. Rather than treating individual analytical outputs independently, the results are interpreted as components of a broader conceptual framework describing the organization of epigenomic research(Griffiths & Steyvers, 2004; Blei, 2012). The analyses indicate that multiple analytical dimensions—including epigenetic mechanisms, chromatin variability expressions, gene–epigenetic relationships, cell-type specificity, disease associations, temporal dynamics, and environmental influences—collectively form an interconnected knowledge architecture within the literature. These components do not function as isolated observations but rather represent recurring conceptual structures that shape the epigenomic research landscape. From this perspective, epigenomic research can be interpreted as a dynamic regulatory system in which biological mechanisms, environmental influences, and temporal processes interact to shape chromatin states and gene regulation patterns. Literature-level semantic network analysis further confirms that these concepts frequently co-occur within shared research contexts, indicating the existence of a stable conceptual framework within the field(Griffiths & Steyvers, 2004; Blei, 2012). The variability-related indicators derived from the literature corpus provide additional support for this interpretation by quantitatively capturing how consistently variability-related concepts are articulated across studies. Together, these findings suggest that the epigenomic literature reflects a structured and dynamically evolving knowledge system rather than a collection of isolated experimental observations.

Table 8. Summary of biological implications and research linkages derived from literature-level text mining. This tble 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**

 doi:10.20944/preprints202604.0029.v1

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**	**Systems-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**

7.2. Key Biomarker Insights from Literature Mining

The literature corpus analysis reveals several consistent patterns in epigenomic biomarker research. Among epigenomic regulatory mechanisms, DNA methylation remains the most widely investigated biomarker, appearing across diverse disease domains such as cancer, metabolic disorders, neurological diseases, and aging-related conditions. Recurring experimental methodologies—including bisulfite sequencing, methylation arrays, and chromatin accessibility assays—form a stable methodological foundation for epigenomic biomarker studies. The widespread adoption of these technologies suggests that epigenomic biomarker research has developed a relatively standardized experimental framework. Topic clustering analysis further indicates that epigenomic biomarker studies frequently intersect with broader research themes, including gene regulation, chromatin organization, disease epigenetics, and environmental exposure(Griffiths & Steyvers, 2004; Blei, 2012). The repeated appearance of specific biomarkers across multiple disease contexts suggests that certain epigenomic signatures may serve as cross-domain molecular indicators, supporting their potential utility in translational health monitoring applications.

 doi:10.20944/preprints202604.0029.v1

7.3. Chromatin Variability as a Dynamic Health Indicator

One of the central conceptual insights emerging from the literature analysis is the importance of chromatin variability and epigenomic dynamics. Rather than representing static biological markers, epigenomic states often reflect dynamic regulatory processes influenced by environmental exposures, aging, metabolic stress, and disease progression (9, 10, 12). The structural characteristics of chromatin variability inferred from DNA sequence context are illustrated in Figure 5(a), which demonstrates how sequence features may influence chromatin accessibility and regulatory activity**(**Griffiths & Steyvers, 2004; Blei, 2012**)**. Additional modeling results highlighting chromatin variability patterns are presented in Figure 5(b) and Figure 5(c), providing evidence that epigenomic variability may reflect underlying regulatory flexibility in genomic regions. In this context, chromatin variability can be interpreted as a dynamic indicator of biological system stability and adaptive capacity. High variability in certain genomic regions may reflect regulatory plasticity, while reduced variability may indicate epigenetic rigidity associated with disease states. Consequently, variability-related epigenomic features may serve as informative signals for monitoring long-term physiological trajectories in digital health monitoring systems.

Literature-level analysis indicates that SNP-related terminology frequently co-occurs with chromatin state and regulatory element annotations, suggesting an emerging research interest in the interaction between genetic variation and epigenomic dynamics.

However, such co-occurrence patterns should be interpreted cautiously, as they do not provide direct evidence of causal or mechanistic relationships between disease-associated SNPs and chromatin variability.

7.4. Literature-Derived Biomarker Prioritization

Large-scale literature analysis provides a systematic approach for identifying candidate biomarkers with high translational potential. Biomarkers that repeatedly appear across multiple studies, disease contexts, and experimental platforms are more likely to represent robust biological signals rather than study-specific artifacts. The keyword co-occurrence patterns and contextual relationships identified through concordance analysis are summarized in Table 1, and Table 2, which present the query patterns and contextual structures used to retrieve epigenomic biomarker-related literature. Text mining methods therefore offer a complementary strategy for literature-driven biomarker prioritization, where frequency patterns, contextual co-occurrence, and temporal stability of epigenomic terms can help identify molecular features suitable for integration into predictive health monitoring models**(**Griffiths & Steyvers, 2004; Blei, 2012**)**. Such literature-derived prioritization frameworks can assist researchers in narrowing the large space of possible epigenomic markers to a smaller set of candidates that are both biologically meaningful and consistently supported by empirical evidence (13, 15, 18).

8. Translating Epigenomic Biomarkers into AI-Driven Health Monitoring

Figure 7 Conceptual architecture of an AI-driven health monitoring system integrating epigenomic biomarkers with wearable sensor data. The framework combines molecular biomarkers such as DNA methylation and chromatin variability with real-time physiological signals obtained from digital health devices. Multimodal AI models integrate these heterogeneous data streams to infer health states, detect emerging risks, and support predictive and personalized health monitoring.

 doi:10.20944/preprints202604.0029.v1

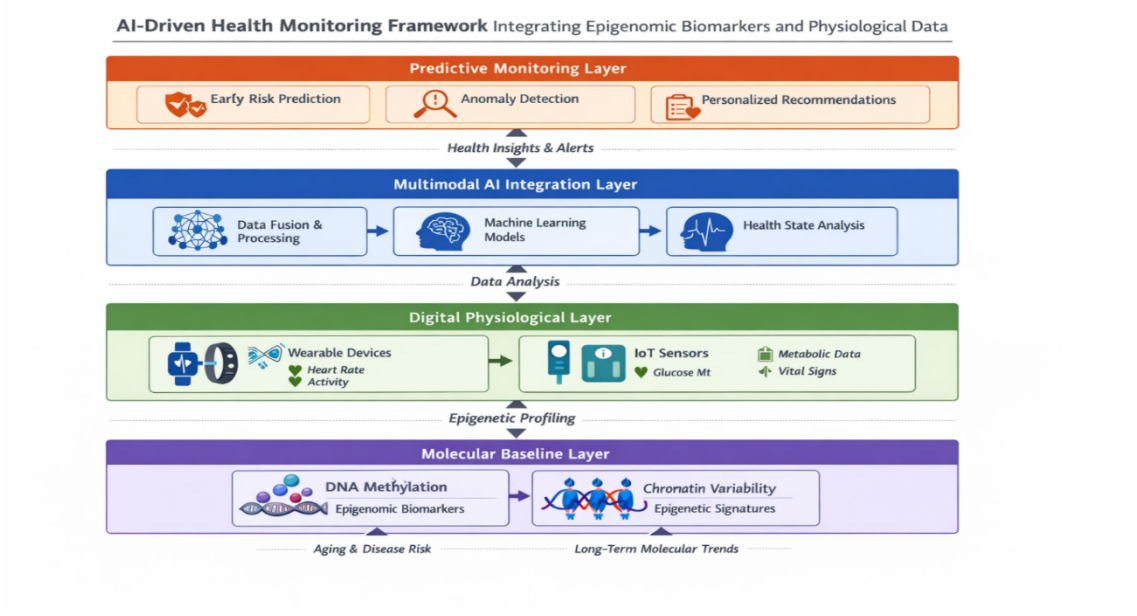


Figure 7. Conceptual architecture of an AI-driven health monitoring system integrating epigenomic biomarkers. This figure presents a multi-layer framework combining molecular biomarkers (e.g., DNA methylation and chromatin variability) with real-time physiological data from wearable devices, integrated through AI models for predictive and personalized health monitoring.

8.1. From Literature-Level Insights to Health Monitoring Architecture

The insights derived from literature-level analysis extend beyond descriptive summaries of epigenomic research trends. By synthesizing large-scale patterns across thousands of publications, the present study identifies recurring experimental configurations, frequently studied epigenomic biomarkers, and stable conceptual relationships between DNA methylation and disease-related biological processes. The overall analytical workflow of the literature mining process is illustrated in **Figure 1**, while the structure of the epigenomic literature corpus and the analytical pipeline are summarized in **Figure 2**. These analyses provide a structured knowledge base that can inform systematic frameworks for biomarker prioritization and translational health monitoring applications. In particular, literature-derived insights can be integrated into a conceptual architecture for **next-generation health monitoring systems**, as illustrated in **Figure 6 and Figure 7**. Within this architecture, epigenomic biomarkers serve as molecular-level indicators that complement physiological signals obtained from wearable devices and clinical monitoring technologies **(**Topol, 2019; Esteva et al., 2019**).**

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

A multi-layer monitoring framework integrating epigenomic biomarkers with digital health technologies may consist of several complementary components.

- **Molecular baseline layer**

Epigenomic biomarkers, including DNA methylation signatures, provide relatively stable indicators of long-term biological states such as aging trajectories, environmental exposure effects, and disease susceptibility (88, 102; Topol, 2019; Esteva et al., 2019).

- **Digital physiological layer**

Wearable devices and IoT-based sensors capture high-frequency physiological signals, including heart rate dynamics, sleep patterns, physical activity levels, and metabolic indicators.

- **Multimodal AI integration layer**

 doi:10.20944/preprints202604.0029.v1

Machine learning models integrate molecular biomarker information with real-time physiological data to infer latent health states and detect emerging risk patterns (Topol, 2019; Esteva et al., 2019).

- **Predictive monitoring layer**

The integrated outputs enable early disease risk prediction, anomaly detection in physiological trends, and personalized health monitoring recommendations (111).

8.3. Multi-Timescale Health Monitoring Through Epigenomic Biomarkers

Within this integrated framework, epigenomic biomarkers function as **molecular anchors** that contextualize rapidly changing physiological signals. By combining relatively stable molecular indicators with high-frequency physiological data, AI-driven monitoring systems can support multi-timescale health assessment (Topol, 2019; Esteva et al., 2019). Such systems allow long-term biological trajectories—such as aging progression and environmental exposure responses—to be interpreted alongside short-term physiological dynamics captured through digital health devices. This integration provides a foundation for next-generation health monitoring platforms that combine molecular biomarkers, wearable sensing technologies, and machine learning–based predictive modeling.

9. Integration for AI-Driven Epigenomic Health Monitoring

9.1. System-Level Architecture of AI-Driven Epigenomic Health Monitoring

Recent advances in digital health technologies have fundamentally shifted healthcare paradigms from episodic, clinic-centered diagnosis toward continuous and personalized health monitoring. Within this emerging landscape, artificial intelligence (AI), wearable sensing technologies, and Internet-of-Things (IoT) platforms enable the real-time acquisition of high-resolution physiological data, including heart rate variability, sleep patterns, physical activity, and metabolic indicators (Miotto et al., 2018; Rajkomar et al., 2018).

Despite these advances, most existing health monitoring systems remain predominantly dependent on short-term physiological signals and lack stable molecular indicators capable of capturing long-term biological states. This limitation constrains their ability to model chronic disease progression, cumulative environmental exposure, and aging-related biological changes.

Epigenomic biomarkers, particularly DNA methylation signatures, provide a critical opportunity to address this gap. Unlike rapidly fluctuating sensor-derived signals, epigenomic patterns evolve over extended timescales and encode long-term biological processes, including aging dynamics, chronic inflammation, environmental exposure, and disease susceptibility. Consequently, they can serve as stable molecular references that complement high-frequency physiological observations.

From a systems perspective, epigenomic biomarkers can be conceptualized as low-frequency biological priors, while wearable sensor data represent high-frequency physiological observations. The integration of these complementary data modalities enables the construction of hierarchical health monitoring systems capable of capturing both long-term biological trajectories and short-term physiological variations.

To operationalize this integration, an AI-driven epigenomic health monitoring system can be structured as a multi-layer architecture:

- **Molecular baseline layer:** This layer establishes individualized biological reference profiles using epigenomic biomarkers derived from DNA methylation data. These profiles may encode biological age acceleration, chromatin variability, and disease-associated epigenetic signatures.
- **Digital physiological layer:** This layer captures continuous physiological data streams from wearable devices and IoT-based sensors, including cardiovascular dynamics, sleep quality, activity patterns, and metabolic indicators.

 doi:10.20944/preprints202604.0029.v1

- **Multimodal integration layer:** In this layer, machine learning models integrate heterogeneous data sources to infer latent health states(Miotto et al., 2018; Rajkomar et al., 2018). Approaches such as deep neural networks, multimodal representation learning, and probabilistic modeling enable the fusion of molecular and physiological data.
- **Predictive monitoring layer:** This final layer translates integrated data into actionable outputs, including early disease risk prediction, anomaly detection, adaptive health recommendations, and clinical decision support.

This hierarchical architecture enables the transition from reactive healthcare toward predictive and preventive health management.

9.2. Data Integration and Multimodal Analytical Strategies

The integration of epigenomic biomarkers with real-time physiological data requires robust multimodal analytical frameworks capable of handling heterogeneous data types with distinct temporal, structural, and statistical properties.

A key challenge in this context lies in the temporal disparity between data modalities. Epigenomic biomarkers represent relatively stable, low-frequency signals that evolve over weeks to years, whereas wearable sensor data capture high-frequency physiological fluctuations at sub-second to daily resolutions. Bridging this temporal gap requires modeling strategies that can align multi-timescale data within a unified analytical framework.

Several computational approaches can be employed to address this challenge:

- Multimodal representation learning: Learning shared latent representations that capture interactions between molecular and physiological signals.
- Bayesian hierarchical modeling: Incorporating epigenomic features as prior distributions that constrain the interpretation of dynamic physiological data.
- Deep learning architectures: Integrating heterogeneous inputs through attention mechanisms, recurrent neural networks, or transformer-based models to capture both temporal dynamics and cross-modal relationships(Miotto et al., 2018; Rajkomar et al., 2018).

Through these approaches, epigenomic biomarkers function as biological anchors, providing contextual stability that enhances the interpretability and robustness of predictive models derived from high-frequency sensor data.

Importantly, such multimodal integration enables hierarchical health inference, where long-term biological risks and short-term physiological deviations are jointly modeled. This capability is particularly valuable for early disease detection, as subtle deviations from individualized baselines can be identified before the onset of overt clinical symptoms.

9.3. Future Directions for Personalized and Predictive Health Monitoring

The integration of epigenomic biomarkers into AI-driven health monitoring systems opens several promising directions for next-generation precision health.

First, large-scale literature mining, as demonstrated in this study, provides a systematic approach for biomarker prioritization. By identifying recurrent epigenetic signatures, stable chromatin variability patterns, and cross-disease biomarker candidates, literature-level analysis enables the construction of curated biomarker panels that are robust across diverse biological contexts.

Second, future systems are expected to evolve toward adaptive and self-improving health monitoring platforms. By continuously updating predictive models using incoming physiological data and periodically recalibrated epigenomic profiles, these systems can dynamically refine individualized health trajectories.

Third, the integration of epigenomic data with digital health technologies will facilitate the development of multi-timescale health modeling frameworks, in which molecular biomarkers capture long-term biological states while wearable devices provide high-resolution monitoring of

 doi:10.20944/preprints202604.0029.v1

short-term physiological dynamics. This dual-resolution approach enables more accurate and personalized risk prediction.

Finally, the incorporation of epigenomic biomarkers into clinical workflows has the potential to transform preventive medicine. Rather than serving solely as diagnostic endpoints, epigenomic features can function as foundational components of predictive health ecosystems, supporting early intervention, risk stratification, and personalized therapeutic strategies.

Although the relationship between disease-associated SNPs and chromatin variability remains to be fully elucidated, integrating these complementary biological signals may provide a more comprehensive representation of individual health states.

In this context, SNPs can be conceptualized as stable genetic baselines, while chromatin variability reflects dynamic regulatory responses, together offering a multi-layered framework for future AI-driven health monitoring systems.

In summary, the convergence of epigenomics, AI, and digital health technologies enables a paradigm shift toward proactive, data-driven healthcare systems. By integrating stable molecular indicators with dynamic physiological data, AI-driven epigenomic monitoring frameworks provide a scalable foundation for the development of next-generation precision health platforms capable of detecting subtle biological changes long before clinical manifestation.

10. Future Directions: Toward Integrative Modeling of SNPs, Chromatin Variability, and AI-Driven Health Monitoring

10.1. Unresolved Relationships Between Genetic Variation and Chromatin Variability

A fundamental unresolved question in epigenomic research concerns the relationship between genetic variation and chromatin state dynamics. While chromatin variability captures the dynamic regulatory behavior of genomic regions across cell types and biological conditions, the extent to which this variability is influenced or constrained by underlying DNA sequence variation remains incompletely understood.

In particular, disease-associated single nucleotide polymorphisms (SNPs) have been widely studied in the context of genome-wide association studies (GWAS), yet their potential role in modulating chromatin variability has not been systematically established. Although emerging evidence suggests that sequence-level features may influence chromatin accessibility and regulatory activity, direct mechanistic links between SNPs and large-scale chromatin variability patterns remain limited.

Future research should aim to determine whether SNPs act as structural constraints, probabilistic modulators, or independent factors in shaping chromatin state transitions. Addressing this question will require integrative analyses that combine genomic variation data with high-resolution epigenomic profiling across diverse biological contexts.

10.2. Integrative Multi-Omics and Multi-Scale Data Modeling

Advancing the understanding of SNP–chromatin interactions necessitates the integration of heterogeneous data types spanning multiple biological scales. Genomic data (e.g., SNPs), epigenomic profiles (e.g., DNA methylation, chromatin accessibility), and transcriptomic outputs collectively define regulatory landscapes that underlie cellular function and disease processes.

One of the key challenges in this domain lies in reconciling differences in temporal and structural resolution across data modalities. Genetic variation represents a largely static component of the genome, whereas chromatin states and epigenomic markers exhibit dynamic and context-dependent behavior. Bridging this gap requires modeling frameworks capable of integrating static genetic constraints with dynamic epigenomic responses.

Machine learning approaches, including deep neural networks, graph-based models, and probabilistic frameworks, offer promising avenues for modeling such complex relationships. In particular, multimodal learning architectures can capture cross-layer dependencies, enabling the

 doi:10.20944/preprints202604.0029.v1

identification of latent representations that link SNPs to chromatin variability and downstream phenotypic outcomes.

10.3. Literature-Level Knowledge Integration for Biomarker Prioritization

The rapid expansion of epigenomic research has generated a vast and heterogeneous body of literature, making systematic knowledge integration increasingly important. Literature-level text mining, as demonstrated in this study, provides a scalable approach for identifying recurring patterns, stable conceptual associations, and emerging research themes.

In the context of SNP–chromatin relationships, literature mining can be leveraged to detect co-occurrence patterns between genetic variation, chromatin states, regulatory elements, and disease phenotypes. While such patterns do not constitute direct evidence of causal relationships, they can highlight promising hypotheses and guide the prioritization of candidate genomic regions for further investigation.

Future work should focus on developing knowledge-guided modeling frameworks, in which insights derived from literature mining are incorporated into computational pipelines for biomarker discovery. Such approaches may enhance the robustness, interpretability, and translational relevance of AI-driven analyses.

10.4. Toward AI-Driven Multi-Timescale Health Monitoring Frameworks

Integrating SNPs and chromatin variability into health monitoring systems offers a conceptual pathway toward multi-timescale modeling of human health. In such a framework, SNPs can be interpreted as stable genetic baselines that define individual-specific regulatory potential, while chromatin variability reflects dynamic epigenomic responses to environmental and physiological factors.

When combined with high-frequency physiological data obtained from wearable devices and IoT-based sensors, these biological layers enable a hierarchical representation of health states across different temporal resolutions. This multi-layer integration allows for the simultaneous modeling of long-term biological trajectories and short-term physiological fluctuations.

However, the practical realization of such systems requires careful validation of the relationships between genetic variation, epigenomic dynamics, and health outcomes. Future studies should therefore aim to establish whether integrating SNPs and chromatin variability into predictive models yields measurable improvements in early disease detection, risk stratification, and personalized health management.

10.5. Toward Hypothesis-Driven and Translational Epigenomic Research

Moving forward, it is essential to transition from descriptive analyses toward hypothesis-driven and translational research frameworks. The potential interaction between SNPs and chromatin variability represents a biologically plausible yet underexplored interface that may provide new insights into gene regulation, disease mechanisms, and biomarker discovery.

To realize this potential, interdisciplinary efforts are required to integrate genomic data, epigenomic profiling, computational modeling, and clinical validation. Large-scale cohort studies, longitudinal data collection, and multimodal data integration will be critical for establishing robust and reproducible findings.

Ultimately, elucidating the interplay between genetic variation and epigenomic dynamics may enable the development of next-generation AI-driven health monitoring systems. Such systems would move beyond reactive diagnostics toward proactive and predictive healthcare, supporting early intervention and personalized medicine at an unprecedented scale.

 doi:10.20944/preprints202604.0029.v1

11. Conclusion

This review reconstructs the landscape of epigenomic and DNA methylation research from a literature-level perspective, providing an integrated view of structural patterns, methodological regularities, and temporal dynamics relevant to biomarker discovery and health monitoring. Using large-scale text mining of full-text corpora, we identified recurring experimental designs, stable conceptual structures, and long-term evolutionary trends across the epigenomic research field. Our analysis shows that epigenomic research develops through multiple parallel domains—including disease-focused studies, chromatin regulatory mechanisms, transcriptome-integrative analyses, and cancer epigenomics—rather than following a single linear trajectory. The recurrence of methodological combinations and the contextual stability of key concepts suggest that the epigenomic literature has evolved through relatively consistent analytical practices and shared research frameworks.

Temporal analysis further indicates rapid expansion of epigenomic research after 2010, with increasing emphasis on variability, heterogeneity, and dynamic regulation. These findings demonstrate the potential of AI-driven literature mining to systematically organize large-scale biomedical knowledge and support biomarker prioritization. By transforming dispersed research findings into structured knowledge architectures, literature-level analysis provides a computational foundation for integrating epigenomic biomarkers into AI-enabled health monitoring systems. The convergence of epigenomic biomarker science, semantic AI, and digital health infrastructures may ultimately support more adaptive and data-driven precision health monitoring frameworks.

Author Contributions: Conceptualization, H.-S.P.; methodology, J.-H.O.; software, J.-H.O.; validation, H.-S.P.; formal analysis, J.-H.O.; 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.; supervision, H.-S.P. All authors have read and agreed to the published version of the manuscript.

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 6,152 full-text articles published before January 27, 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-literature-mining. Additional scripts and analysis tools used in this study are available from the authors upon reasonable request.

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

Appendix A. Glossary

Table A1. *Key terms used in this study.*

Term	Description
ATAC-seq	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.

 doi:10.20944/preprints202604.0029.v1

(Assay for Transposase-Accessible Chromatin using sequencing)	
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 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.

 doi:10.20944/preprints202604.0029.v1

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. Jones, P.A.; Baylin, S.B. The fundamental role of epigenetic events in cancer. Nat. Rev. Genet. 2002, 3, 415–428.
2. Bird, A. Perceptions of epigenetics. Nature 2007, 447, 396–398.
3. Portela, A.; Esteller, M. Epigenetic modifications and human disease. Nat. Biotechnol. 2010, 28, 1057–1068.
4. Lim, I.; Tan, J.; Alam, A.; Idrees, M.; Brenan, P.A.; Coletta, R.D.; Kujan, O. Epigenetics in the diagnosis and prognosis of head and neck cancer: A systematic review. J. Oral Pathol. Med. 2024, 53, 90–106.
5. Burkitt, K. Role of DNA methylation profiles as potential biomarkers and novel therapeutic targets in head and neck cancer. Cancers 2023, 15, 4685. https://doi.org/10.3390/cancers15194685
6. Villicaña, S.; Castillo-Fernandez, J.; Hannon, E.; Christiansen, C.; Tsai, P.-C.; Maddock, J.; Kuh, D.; Suderman, M.; Power, C.; Relton, C.; et al. Genetic impacts on DNA methylation help elucidate regulatory genomic processes. Genome Biol. 2023, 24, 11. https://doi.org/10.1186/s13059-023-03011
7. Feehley, T.; O'Donnell, C.W.; Mendlein, J.; Karande, M.; McCauley, T. Drugging the epigenome in the age of precision medicine. Clin. Epigenetics 2023, 15, 6.
8. Nadiger, N.; Veed, J.K.; Chinya Nataraj, P.; Mukhopadhyay, A. DNA methylation and type 2 diabetes: A systematic review. Clin. Epigenetics 2024, 16, 67.
9. 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.
10. 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.
11. Bao-Caamano, A.; Costa-Fraga, N.; Cayrefourcq, L.; Jácome, M.A.; Rodriguez-Casanova, A.; Muinelo-Romay, L.; Díaz-Lagares, A. Epigenomic analysis reveals a unique DNA methylation program of metastasis-competent circulating tumor cells in colorectal cancer. Sci. Rep. 2023, 13, 15401.
12. Novoa, J.; Chagoyen, M.; Benito, C.; Moreno, F.J.; Pazos, F. Pmidigest: Interactive review of large collections of PubMed entries to distill relevant information. Genes 2023, 14, 942.
13. Smalheiser, N.R.; Fragnito, D.P.; Tirk, E.E. Anne O'Tate: Value-added PubMed search engine for analysis and text mining. PLoS ONE 2021, 16, e0248335.
14. Zhao, S.; Su, C.; Lu, Z.; Wang, F. Recent advances in biomedical literature mining. Brief. Bioinform. 2021, 22, bbaa057.
15. Ye, Z.; Tafti, A.P.; He, K.Y.; Wang, K.; He, M.M. Sparktext: Biomedical text mining on big data framework. PLoS ONE 2016, 11, e0162721.
16. Comeau, D.C.; Wei, C.H.; Dogan, R.I.; Lu, Z. PMC text mining subset in BioC: 2.3 million full text articles and growing. arXiv 2018, arXiv:1804.05957.

 doi:10.20944/preprints202604.0029.v1

17. Westergaard, D.; Stærfeldt, H.H.; Tønsberg, C.; Jensen, L.J.; Brunak, S. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. PLoS Comput. Biol. 2018, 14, e1005962.
18. Cohen, K.B.; Johnson, H.L.; Verspoor, K.; Roeder, C.; Hunter, L.E. The structural and content aspects of abstracts versus bodies of full text journal articles are different. BMC Bioinform. 2010, 11, 492.
19. Tudor, C.O.; Ross, K.E.; Li, G.; Vijay-Shanker, K.; Wu, C.H.; Arighi, C.N. Construction of phosphorylation interaction networks by text mining of full-length articles using the eFIP system. Database 2015, 2015, bav020.
20. Islamaj Doğan, R.; Kim, S.; Chatr-Aryamontri, A.; Chang, C.S.; Oughtred, R.; Rust, J.; Tyers, M. The BioC-BioGRID corpus: Full text articles annotated for curation of protein–protein and genetic interactions. Database 2017, 2017, baw147.
21. Aria, M.; Cuccurullo, C. bibliometrix: An R-tool for comprehensive science mapping analysis. J. Informetr. 2017, 11, 959–975.
22. Donthu, N.; Kumar, S.; Mukherjee, D.; Pandey, N.; Lim, W.M. How to conduct a bibliometric analysis: An overview and guidelines. J. Bus. Res. 2021, 133, 285–296.
23. Chen, C. Science mapping: A systematic review of the literature. J. Data Inf. Sci. 2017, 2, 1–40.
24. Van Eck, N.J.; Waltman, L. Visualizing bibliometric networks. In Measuring Scholarly Impact: Methods and Practice; Springer: Cham, Switzerland, 2014; pp. 285–320.
25. Börner, K.; Chen, C.; Boyack, K.W. Visualizing knowledge domains. Annu. Rev. Inf. Sci. Technol. 2003, 37, 179–255.
26. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet allocation. J. Mach. Learn. Res. 2003, 3, 993–1022.
27. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proc. NAACL-HLT; 2019; pp. 4171–4186.
28. Wei, C.H.; Kao, H.Y.; Lu, Z. PubTator: A web-based text mining tool for assisting biocuration. Nucleic Acids Res. 2013, 41, W518–W522.
29. Johnson, D.S.; Mortazavi, A.; Myers, R.M.; Wold, B. Genome-wide mapping of in vivo protein–DNA interactions. Science 2007, 316, 1497–1502.
30. Comeau, D.C.; Wei, C.H.; Islamaj Doğan, R.; Lu, Z. PMC text mining subset in BioC: About three million full-text articles and growing. Bioinformatics 2019, 35, 3533–3535.
31. Horvath, S. DNA methylation age of human tissues and cell types. Genome Biol. 2013, 14, R115.
32. Hannum, G.; Guinney, J.; Zhao, L.; Zhang, L.; Hughes, G.; Sadda, S.; Zhang, K. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol. Cell 2013, 49, 359–367.
33. Levine, M.E.; Lu, A.T.; Quach, A.; Chen, B.H.; Assimes, T.L.; Bandinelli, S.; Horvath, S. An epigenetic biomarker of aging for lifespan and healthspan. Aging 2018, 10, 573–591.
34. Lu, A.T.; Quach, A.; Wilson, J.G.; Reiner, A.P.; Aviv, A.; Raj, K.; Horvath, S. DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging 2019, 11, 303–327.
35. 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.
36. 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.
37. 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.
38. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model. Bioinformatics 2020, 36, 1234–1240.
39. Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Poon, H. Domain-specific language model pretraining. ACM Trans. Comput. Healthc. 2021, 3, 1–23.
40. Xu, Z. DNA methylation-based health predictors. Epigenomics 2025, 17, 1083–1090.
41. Kiselev, I.S.; Baulina, N.M.; Favorova, O.O. Epigenetic clock. Biochemistry (Moscow) 2025, 90, S356–S372.
42. Chen, Y.; Cheng, X.; Ji, S. DNA methylation and prediction of biological age. Front. Mol. Biosci. 2025, 12, 1734464.
43. Horvath, S.; Raj, K. DNA methylation-based biomarkers and aging. Nat. Rev. Genet. 2018, 19, 371–384.

 doi:10.20944/preprints202604.0029.v1

44. Martínez-Iglesias, O.; Naidoo, V.; Corzo, L.; Pego, R.; Seoane, S.; Rodríguez, S.; Cacabelos, R. DNA methylation as a biomarker for disease outcome. Genes 2023, 14, 365.
45. Janovska, J.; Nixdorff, U.N.; Voicehovska, J.V. DNA methylation as a biomarker for cardiometabolic risk. Eur. J. Prev. Cardiol. 2023, 30, zwad125.
46. Kim, H.; Wang, X.; Jin, P. Developing DNA methylation-based diagnostic biomarkers. J. Genet. Genomics 2018, 45, 87–97.
47. Sahoo, K.; Lingasamy, P.; Khatun, M.; Sudhakaran, S.L.; Salumets, A.; Sundararajan, V.; Modhukur, V. Artificial intelligence in cancer epigenomics. Epigenetics Chromatin 2025, 18, 35.
48. Levy, J.J.; Diallo, A.B.; Saldias Montivero, M.K.; Gabbita, S.; Salas, L.A.; Christensen, B.C. Aging prediction with AI-based epigenetic clocks. Epigenomics 2025, 17, 49–57.
49. Kalyakulina, A.; Yusipov, I.; Trukhanov, A.; Franceschi, C.; Moskalev, A.; Ivanchenko, M. EpInflammAge. Int. J. Mol. Sci. 2025, 26, 6284.
50. 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.
51. 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.
52. Feinberg, A.P. Phenotypic plasticity and the epigenetics of human disease. Nature 2007, 447, 433–440.
53. Baylin, S.B.; Jones, P.A. A decade of exploring the cancer epigenome. Nat. Rev. Cancer 2011, 11, 726–734.
54. Esteller, M. Epigenetics in cancer. N. Engl. J. Med. 2008, 358, 1148–1159.
55. Ioannidis, J.P.A. Why most clinical research is not useful. PLoS Med. 2016, 13, e1002049.
56. Ernst, J.; Kellis, M. ChromHMM: automating chromatin-state discovery. Nat. Methods 2012, 9, 215–216.
57. Kundaje, A.; Meuleman, W.; Ernst, J.; et al. Integrative analysis of 111 reference human epigenomes. Nature 2015, 518, 317–330.
58. Hearst, M.A. Untangling text data mining. ACL 1999.
59. Feldman, R.; Sanger, J. The Text Mining Handbook; Cambridge University Press: Cambridge, UK, 2007.
60. Khabsa, M.; Giles, C.L. The number of scholarly documents on the public web. PLoS ONE 2014, 9, e93949.
61. West, J.D.; Jacquet, J.; King, M.M.; Correll, S.J.; Bergstrom, C.T. The role of text in scientific article analysis. J. Informetrics 2013, 7, 487–499.
62. Griffiths, T.L.; Steyvers, M. Finding scientific topics. Proc. Natl. Acad. Sci. USA 2004, 101 (Suppl. 1), 5228–5235.
63. Blei, D.M. Probabilistic topic models. Commun. ACM 2012, 55, 77–84.
64. Topol, E.J. High-performance medicine. Appl. Sci. 2019, 25, 44–56.
65. 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.
66. Miotto, R.; Li, L.; Kidd, B.A.; Dudley, J.T. Deep learning for healthcare. Appl. Sci. 2018, 19, 1236–1246.
67. Rajkomar, A.; Dean, J.; Kohane, I. Scalable deep learning with EHR. Appl. Sci. 2018, 1, 18.
68. Ioannidis, J.P.A. Why most published research findings are false. PLoS Med. 2005, 2, e124.
69. Beam, A.L.; Kohane, I.S. Big data and machine learning in healthcare. JAMA 2018, 319, 1317–1318.