DATABASE

Open Access

Data stewardship and curation practices in AI-based genomics and automated microscopy image analysis for high-throughput screening studies: promoting robust and ethical AI applications

Asefa Adimasu Taddese[1], Assefa Chekole Addis[2] and Bjorn T. Tam[1,3*]

Abstract

Background Researchers have increasingly adopted AI and next-generation sequencing (NGS), revolutionizing genomics and high-throughput screening (HTS), and transforming our understanding of cellular processes and disease mechanisms. However, these advancements generate vast datasets requiring effective data stewardship and curation practices to maintain data integrity, privacy, and accessibility. This review consolidates existing knowledge on key aspects, including data governance, quality management, privacy measures, ownership, access control, accountability, traceability, curation frameworks, and storage systems.

Methods We conducted a systematic literature search up to January 10, 2024, across PubMed, MEDLINE, EMBASE, Scopus, and additional scholarly platforms to examine recent advances and challenges in managing the vast and complex datasets generated by these technologies. Our search strategy employed structured keyword queries focused on four key thematic areas: data governance and management, curation frameworks, algorithmic bias and fairness, and data storage, all within the context of AI applications in genomics and microscopy. Using a realist synthesis methodology, we integrated insights from diverse frameworks to explore the multifaceted challenges associated with data stewardship in these domains. Three independent reviewers, who systematically categorized the information across critical themes, including data governance, quality management, security, privacy, ownership, and access control conducted data extraction and analysis. The study also examined specific AI considerations, such as algorithmic bias, model explainability, and the application of advanced cryptographic techniques. The review process included six stages, starting with an extensive search across multiple research databases, resulting in 273 documents. Screening based on broad criteria, titles, abstracts, and full texts followed this, narrowing the pool to 38 highly relevant citations.

Results Our findings indicated that significant research was conducted in 2023 by highlighting the increasing recognition of robust data governance frameworks in AI-driven genomics and microscopy. While 36 articles extensively discussed data interoperability and sharing, AI-model explain ability and data augmentation remained underexplored, indicating significant gaps. The integration of diverse data types—ranging from sequencing and clinical

*Correspondence:
Bjorn T. Tam
bjornttam@gmail.com
Full list of author information is available at the end of the article

data to proteomic and imaging data—highlighted the complexity and expansive scope of AI applications in these fields. The current challenges identified in AI-based data stewardship and curation practices are lack of infrastructure and cost optimization, ethical and privacy considerations, access control and sharing mechanisms, large scale data handling and analysis and transparent data-sharing policies and practice. Proposed solutions to address issues related to data quality, privacy, and bias management include advanced cryptographic techniques, federated learning, and blockchain technology. Robust data governance measures, such as GA4GH standards, DUO versioning, and attribute-based access control, are essential for ensuring data integrity, security, and ethical use. The study also emphasized the critical role of Data Management Plans (DMPs), meticulous metadata curation, and advanced cryptographic techniques in mitigating risks related to data security and identifiability. Despite advancements, significant challenges persisted in balancing data ownership with research accessibility, integrating heterogeneous data sources, ensuring platform interoperability, and maintaining data quality. Ongoing risks of unauthorized access and data breaches underscored the need for continuous innovation in data management practices and stricter adherence to legal and ethical standards.

Conclusions These findings explored the current practices and challenges in data stewardship, offering a roadmap for strengthening the governance, security, and ethical use of AI in genomics and microscopy. While robust governance frameworks and ethical practices have established a foundation for data integrity and transparency, there remains an urgent need for collaborative efforts to develop interoperable platforms and transparent data-sharing policies. Additionally, evolving legal and ethical frameworks will be crucial to addressing emerging challenges posed by AI technologies. Fostering transparency, accountability, and ethical responsibility within the research community will be key to ensuring trust and driving ethically sound scientific advancements.

Keywords Data stewardship, Data curation, Artificial intelligence, Genomics, Microscopy image analysis, Scoping review

Introduction

Currently, advancements in high-throughput technologies, determined by Artificial Intelligence (AI) and next-generation sequencing (NGS), have changed the trends of genomic research and high-throughput screening (HTS) studies [1–3]. Alongside, the emergence of complex high-content screening (HCS), single-cell omics technologies and automated microscopy image acquisition and analysis, has become paramount for understanding the molecular and cellular processes and facilitating drug discovery in HTS studies [4]. These cutting-edge technologies have enabled the interoperable big datasets, spanning genomics, proteomics, microbiomics, and radiomics, thereby shedding light on intricate cellular complexities and disease mechanisms [5–9]. For instance, single-cell Sequencing techniques provide insights into cellular heterogeneity, while deep visual proteomics identifies disease-associated protein markers with precision and depth [2]. Genomic data, such as single-cell sequencing, enables the mapping of cell types and states, revealing cellular complexity and heterogeneity [10]. Proteomic data obtained through deep visual proteomics allows for unbiased characterization of cellular function and identification of protein markers associated with specific phenotypes and disease states [11, 12]. Furthermore, microscopy image analysis techniques, such as Cell Painting, equip high-throughput data decisive for predicting drug activity and toxicity, particularly in HTS studies [13]. These techniques also provide insightful information about cellular function, disease mechanisms, and treatment approaches, which aids in the development of novel, life-saving treatments [14–16]. The multidisciplinary nature of this field is demonstrated in personalized oncology and precision medicine, where the combination of various data types with in-vitro drug sensitivity and resistance tests (DSRT) informs personalized therapies [2], as well as in drug discovery, where HCS advances unbiased compound screening [17].

Despite its potential to revolutionize high-throughput screening, the sheer volume and complexity of data generated by these technologies pose a substantial challenge to data stewardship and curation, particularly in the context of AI-based genomics and automated microscope image analysis for HTS research [18–21]. Furthermore, incorporating AI-based techniques improves the analysis of HTS data, emphasizing the importance of strong data management and curation practices [22–25].

While current research efforts document various data management methodologies and technological advancements, a comprehensive understanding of the current landscape, challenges, and emerging trends in data stewardship and curation practices within this specific domain remains limited. Existing literature often focuses on individual aspects or technologies, hindering a holistic understanding of data management strategies. Hence, a systematic scoping review is reasonable to explore the existing literature, identify gaps, and provide insights to inform future research and enhance data management practices in this rapidly evolving field. Therefore, this scoping review aimed to map and synthesize existing literature on Data stewardship and curation practice in AI-based genomics and automated microscopy image analysis for high-throughput screening research in a systematic way. It seeks to identify knowledge gaps, provide a comprehensive overview, and address these gaps to guide future research courses, refine data management practices, and ultimately facilitate more effective and streamlined HTS research endeavors.

Research questions and framing frameworks

The review questions were designed to explore various dimensions of data stewardship, including data governance, quality management, privacy and security, ownership and access control, accountability and traceability, curation frameworks, interoperability, sharing practices, databases, and storage systems. By integrating multiple framing frameworks—namely MIBI (Minimum Information in Biological Imaging), MIAME (Minimum Information About a Microarray Experiment) [26], CASPE (Critical Appraisal Skills Programme for EBM), and FAIRsFAIR (FAIR data infrastructure for FAIR data) [27]—this review ensures a comprehensive investigation. These frameworks offer valuable insights into the responsible management of data in the context of AI-driven genomics and microscopy image analysis research (see Table 1).

Methods

Literature search strategy

A systematic article search was undertaken across four key databases: PubMed, MEDLINE, EMBASE, Scopus and other websites including google scholar, semantic Scholar and Google. These were selected based on their multidisciplinary focus spanning medicine, biotechnology, health data science and governance literature. Structured keyword search strings were built combining Medical Subject Headings terms and free text keywords clustered across four key themes: (i) Data Governance OR Data Quality and Management OR Data Security and Privacy OR Data Ownership and Access Control OR Data Accountability and Traceability AND AI AND Genomics OR High-Throughput Screening OR Automated Microscopy Image Analysis; (ii) Data Curation Frameworks and Tools OR Specific Curation Frameworks OR Curation Tools and Platforms OR Interoperability and Sharing AND AI AND Genomics OR High-Throughput Screening OR Automated Microscopy Image Analysis; (iii) Algorithmic Bias and Fairness OR Explainability and Interpretability OR Data Augmentation and Synthetic Data Generation AND AI AND Genomics OR High-Throughput Screening OR Automated Microscopy Image Analysis and (iv) Data Storage OR Databases AND AI AND Genomics OR High-Throughput Screening OR Automated Microscopy Image Analysis.

Targeted searches for each theme were executed and subsequently pooled. Results were limited to English language publications in peer-reviewed journals within the

Table 1 Research questions and analytical frameworks for scoping review on data stewardship and curation practices in AI-driven genomics and automated microscopy

Specific Research Questions	Framework
RQ1: What is the prevalent data stewardship and curation strategies in AI-driven genomics and microscopy image analysis, and how do they enhance the management and quality control of large datasets??	MIBI (Minimum Information in Biological Imaging)
RQ2: What curation frameworks and platforms facilitate systematic data management in AI-driven genomics and microscopy research?	
RQ3: Which databases and storage systems are most effective for managing genomic and biological data in AI-driven genomics and microscopy research?	
RQ4: How do diverse data types, such as genomic, proteomic, and microbiome data, drive advancements in AI-driven genomics and microscopy image analysis?	MIAME (Minimum Information About a Microarray Experiment)
RQ5: How do data security and privacy measures ensure the protection of sensitive information in AI-driven genomics and microscopy research?	
RQ6: What frameworks regulate data ownership and access control in AI-driven genomics 9and microscopy research?	CASPE (Critical Appraisal Skills Programme for EBM)
RQ7: What are the key challenges in data stewardship within AI-based research, and how can these be addressed to ensure responsible data management?	
RQ8: How do accountability and traceability measures enhance responsible data handling and transparency in AI-driven genomics and microscopy research?	FAIRsFAIR (FAIR data infrastructure for FAIR data)
RQ9: How do interoperability and sharing initiatives promote data accessibility and collaboration in AI-driven genomics and microscopy research?	

past decade. Criteria included consideration of at least one of the four themes, full-text publication, and availability of full-texts for relevance assessment. Abstracts and preprints were excluded. Identified studies were then charted and analyzed to extract relevant information aligned with defined themes. An iterative approach was employed, refining the search strategy and understanding based on emerging insights, feedback, and additional literature identified through snowballing techniques.

Rationale for realist synthesis methodology

Realist synthesis provides a suitable approach for investigating multifaceted, context-dependent topics involving complex interventions or phenomena, aligned to our multi-framework research perspective encompassing varied data stewardship technologies, policies, and practices [28]. Additionally, the integration of diverse question framing frameworks in this review like MIBI, MIAME, CASPE, and FAIRsFAIR necessitates an approach that accommodates varied perspectives and theoretical underpinnings. By adopting a realist synthesis approach, this review aims to elucidate the contextual intricacies influencing the implementation and effectiveness of these interventions.

Data extraction, classification, and analysis

Three independent reviewers (AAT, ACA, TKBT) systematically extracted relevant details from the included articles into a standardized template under the following categories: Study source & setting, research questions/objectives, study methodology, sample/data characteristics, data stewardship strategies, technological tools, limitations and key findings of each article.

Extracted information was iteratively classified across themes. Types of data analyzed or discussed were identified to understand the specific data domains investigated. Specific data stewardship challenges addressed by authors were also extracted to recognize key issues in data management. Data governance, quality management, privacy, security, ownership, access control, and accountability measures were extracted to elucidate strategies for effective data management. Furthermore, data curation frameworks, data interoperability, data sharing measures, algorithmic bias, algorithmic fairness, AI-model explainability, visualization techniques, data augmentation, and data storage systems used were identified.

The analysis and synthesis processes involved a thorough examination of included studies to identify key insights regarding research data stewardship and curation practices reported in AI-based genomics and microscopy image analysis for high-throughput screening. Constructs analyzed encompassed data management, data governance, data security, and data interoperability, alongside AI-specific considerations such as algorithmic bias and model explainability. Reviewers meticulously reviewed each study, extracting relevant data and insights related to the identified constructs. Thematic analysis was then applied to categorize extracted data into thematic groupings based on shared characteristics or concepts. Through iterative refinement and consensus-building exercises, recurring themes were identified and synthesized into a coherent narrative.

The synthesized results were presented using graphs, tables, and narrative summaries, facilitating the interpretation and dissemination of key insights and trends. This approach enabled stakeholders to derive actionable insights into data stewardship and curation practices in AI-based genomics and microscopy image analysis for high-throughput screening.

Results

Stages of the document selection process

Stage 1: Initial Search: The initial stage of the review was conducting a wide-ranging search across four electronic research databases, such as PubMed, MEDLINE, Scopus, and EMBASE. We also used websites including Google Scholar, Google, and Semantic Scholar. The search strategy was focused on identifying articles that addressed specific criteria related to data governance, data quality and management, data security, and privacy, data ownership and access control, and data accountability and traceability within the context of AI applications in genomics, high-throughput screening, or automated microscopy image analysis. A total of 273 documents were retrieved from these databases and websites, forming the initial pool of potential sources for the scoping review.

Stage 2: Number of Citations based on the Initial Broad Criteria: In the initial stage of the search process, citations were obtained based on the predefined broad criteria. The search yielded a total of 96 citations related to data governance, data quality and management, data security and privacy, data ownership and access control, and data accountability and traceability. Additionally, 82 citations were obtained for data curation frameworks and tools, interoperability, and sharing. Furthermore, 25 citations were identified for algorithmic bias and fairness, model explainability and interpretability, and data augmentation and synthetic data generation. Finally, 70 citations were retrieved for data storage and databases in the context of AI and genomics, high-throughput screening, or automated microscopy image analysis.

Stage 3: Screening Citations with Title and Abstract: The third stage involved screening the citations based on their titles and abstracts to assess their relevance to the study criteria. This process aimed at identifying citations that were potentially suitable for inclusion in the review. Following the initial retrieval of citations, subsequent stages of screening were undertaken to refine the selection and retention of relevant documents. During the screening of titles and abstracts, 28 citations were retained for data governance and stewardship, while 35 citations were deemed relevant for data curation frameworks and tools. Additionally, 13 citations were selected for AI-specific considerations, and 43 citations were identified for data storage and databases.

Stage 4: Citations after Second Full Text Screening: At this stage, the citations that passed the title and abstract screening underwent a second round of evaluation through full-text screening. This involved a detailed assessment of the full text of each document to determine its suitability for inclusion in the review. Citations were excluded if they did not meet the study criteria or if they lacked sufficient data to contribute meaningfully to the review. After the second round of screening, which involved a detailed examination of the full text of each citation, the number of retained citations was further reduced. Specifically, 8 citations remained for data governance and stewardship, 6 citations for data curation frameworks and tools, 3 citations for AI-specific considerations, and 7 citations for data storage and databases.

Stage 5: Inclusion of Citations from additional website: Lastly, additional citations were identified through snowballing approaches and direct searching of titles on external platforms. This process yielded 3 citations for data governance and stewardship, 4 citations for data curation frameworks and tools, 2 citations for AI-specific considerations, and 5 citations for data storage and databases.

Stage 6: Citations contribute to the synthesis: Overall, a total of 38 citations were included of which 11 citations were for data governance and stewardship, 10 citations for data curation frameworks and tools, 5 citations for AI-specific considerations, and 12 citations for data storage and databases, contributing to a comprehensive synthesis of evidence in the systematic review (See Fig. 1).

Number of publications by year

The patterns of publications published throughout the years illustrate the different levels of research outputs. There was just one publication each in the years 1999, 2004, 2005, and 2006. A considerable rise in recent research efforts was indicated by the high number of publications recorded in 2021 and the noticeable spike of three and seven publications in 2023 (See Fig. 2).

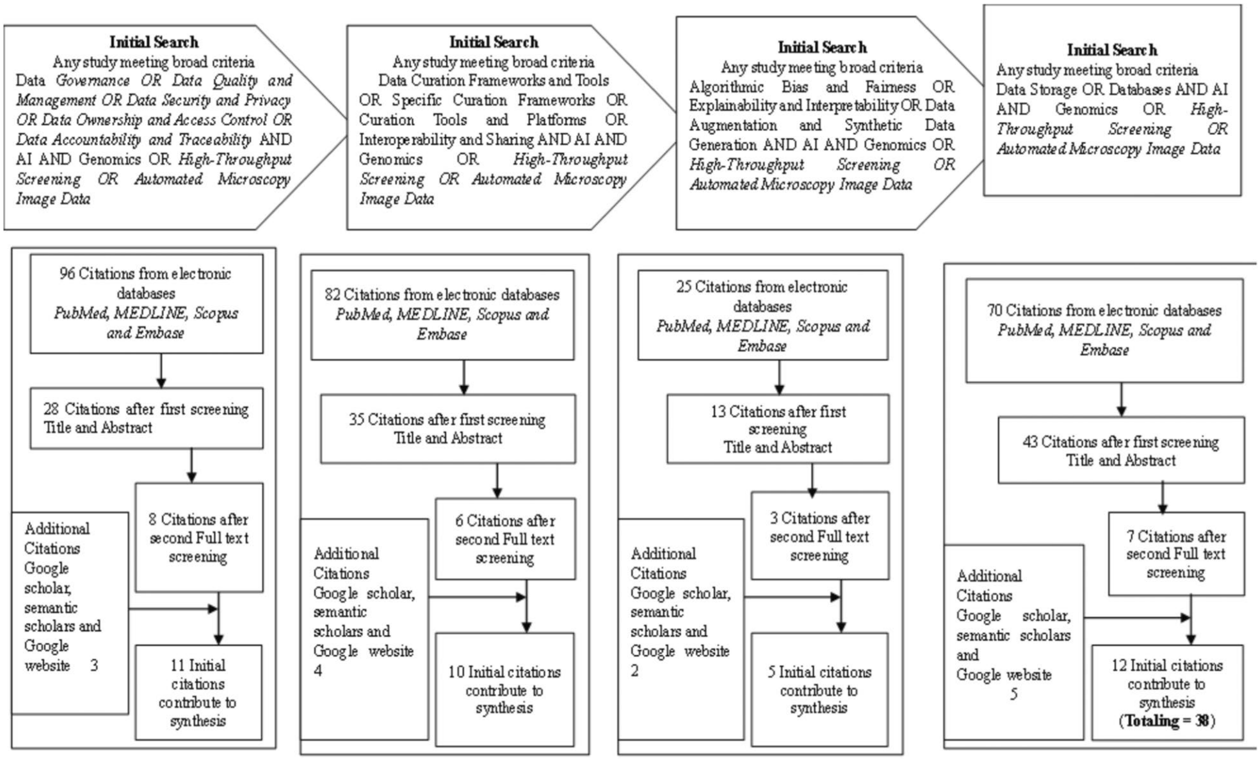

Initial Search
Any study meeting broad criteria
Data Governance OR Data Quality and Management OR Data Security and Privacy OR Data Ownership and Access Control OR Data Accountability and Traceability AND AI AND Genomics OR High-Throughput Screening OR Automated Microscopy Image Data
Initial Search
Any study meeting broad criteria
Data Curation Frameworks and Tools OR Specific Curation Frameworks OR Curation Tools and Platforms OR Interoperability and Sharing AND AI AND Genomics OR High-Throughput Screening OR Automated Microscopy Image Data
Initial Search
Any study meeting broad criteria
Algorithmic Bias and Fairness OR Explainability and Interpretability OR Data Augmentation and Synthetic Data Generation AND AI AND Genomics OR High-Throughput Screening OR Automated Microscopy Image Data
Initial Search
Any study meeting broad criteria
Data Storage OR Databases AND AI AND Genomics OR High-Throughput Screening OR Automated Microscopy Image Data
96 Citations from electronic databases PubMed, MEDLINE, Scopus and Embase
28 Citations after first screening Title and Abstract
8 Citations after second Full text screening
Additional Citations Google scholar, semantic scholars and Google website 3
11 Initial citations contribute to synthesis
82 Citations from electronic databases PubMed, MEDLINE, Scopus and Embase
35 Citations after first screening Title and Abstract
6 Citations after second Full text screening
Additional Citations Google scholar, semantic scholars and Google website 4
10 Initial citations contribute to synthesis
25 Citations from electronic databases PubMed, MEDLINE, Scopus and Embase
13 Citations after first screening Title and Abstract
3 Citations after second Full text screening
Additional Citations Google scholar, semantic scholars and Google website 2
5 Initial citations contribute to synthesis
70 Citations from electronic databases PubMed, MEDLINE, Scopus and Embase
43 Citations after first screening Title and Abstract
7 Citations after second Full text screening
Additional Citations Google scholar, semantic scholars and Google website 5
12 Initial citations contribute to synthesis (Totaling = 38)

Fig. 1 Flow diagram of the search process and article selection in the scoping review

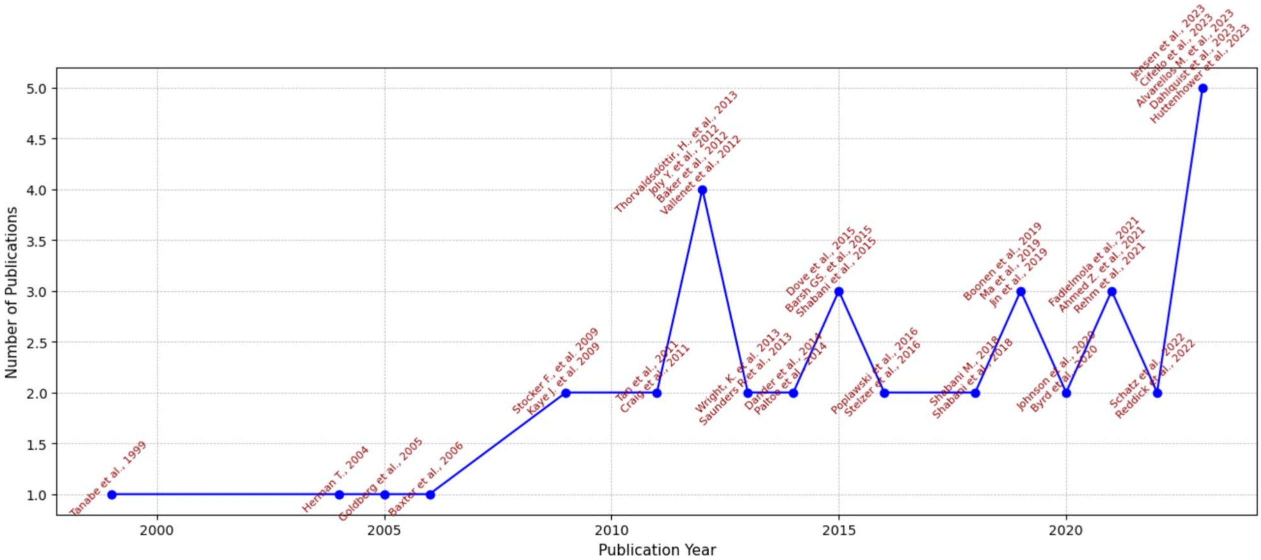


Fig. 2 Annual distribution of publications included in the scoping review

Distribution of publications by data stewardship dimension

The analysis demonstrated varying coverage across the various research data stewardship components, identifying areas of strength and possible gaps: Around 36 articles discussed about data interoperability and sharing measures, and data curation frameworks were reported in 34 articles. Data governance measures, data quality and management measures, and data storage systems were addressed in 32 articles. In contrast, data privacy and security measures and data accountability and traceability measures were discussed in 28 articles. Model explainability and data augmentation or synthetic data management have been addressed in a few studies and require serious attention (See Fig. 3).

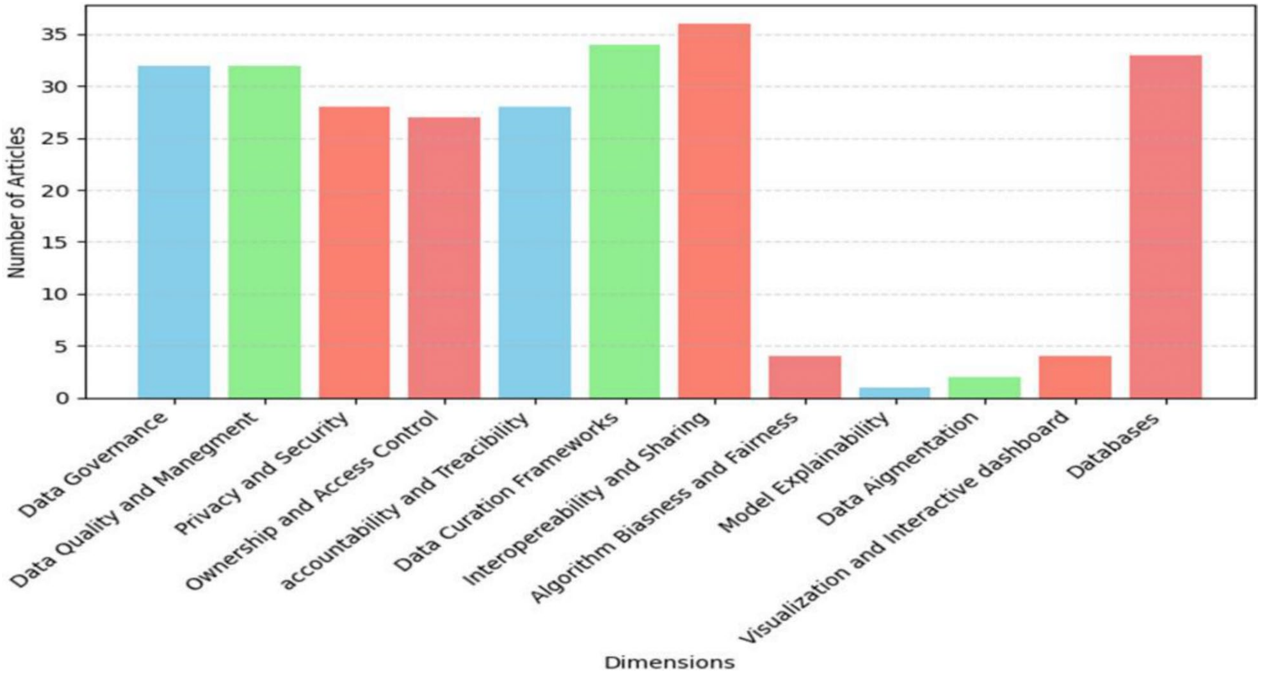


Fig. 3 Distribution of articles addressing data stewardship and curation practices across different dimensions

Datasets and research scope in genomics and microscopy analysis

Researchers have analyzed a wide range of datasets in genomics and microscopy. Schatz et al [29] addressed a wide range of sequencing data, including human genomes, microbiomes, and metagenomes, meanwhile Fadlelmola et al [30] focused on genomic and phenotypic data. Clinical and health data was examined by Wright et al [31] from patients with rare diseases, Thorvaldsdóttir et al [32] from breast cancer patients, and Edward et al from Alzheimer's disease patients, with Rehm et al studying patients with genetic disorders. Ahmed Z. et al investigated proteome and mass spectrometry data from human saliva samples, whereas Stocker, Fischer et al and Ma et al examined experimental and analytical data from yeast and bacterial cells, respectively. Barsh GS. et al and Shabani M investigated legal, ethical, and social data, whereas Tavani and Reddick et al looked at demographic and geographical data. Goldberg et al and Huttenhower et al used imaging and quantitative data to do high-content screening of mouse embryos and image analysis of human microbiomes. Additionally, Saunders R. et al [33] documented experimental protocols and laboratory data to facilitate scientific reproducibility, while Jin et al [34] investigated the governance of blockchain and decentralized data in genomics research (See Fig. 4).

Specific challenges addressed in each article

In addition, our study suggests that the most significant challenges of AI-based data stewardship are lack of infrastructure and cost optimization, ethical and privacy considerations, access control mechanisms, and transparent data-sharing policies. To tackle issues such as data quality, privacy and bias management, advanced cryptographic techniques, federated learning or blockchain technology were suggested. We further discovered that rigorous data governance requirements including standards from GA4GH standards[29], DUO versioning, and attribute-based access control [35], were necessary to maintain integrity of the study data while also ensuring security and ethical use. Data Management Plans (DMPs) [30], extensive metadata curation, and the use of robust cryptographic methodology were identified as critical in addressing data security and identifiability risks.

Schatz et al emphasize the importance of optimizing infrastructure and reducing costs by highlighting the redundancies and inefficiencies in traditional genomics analysis. The complexities of data collection and standardization are highlighted by Fadlelmola et al, especially in African research communities with language differences and sensitive data. The authors Thorvaldsdóttir, H., et al [36] highlight the significance of efficient data management and visualization, especially when dealing with the extensive datasets that are inherent in genomics research. Stocker, Fischer, et al delve into the challenges

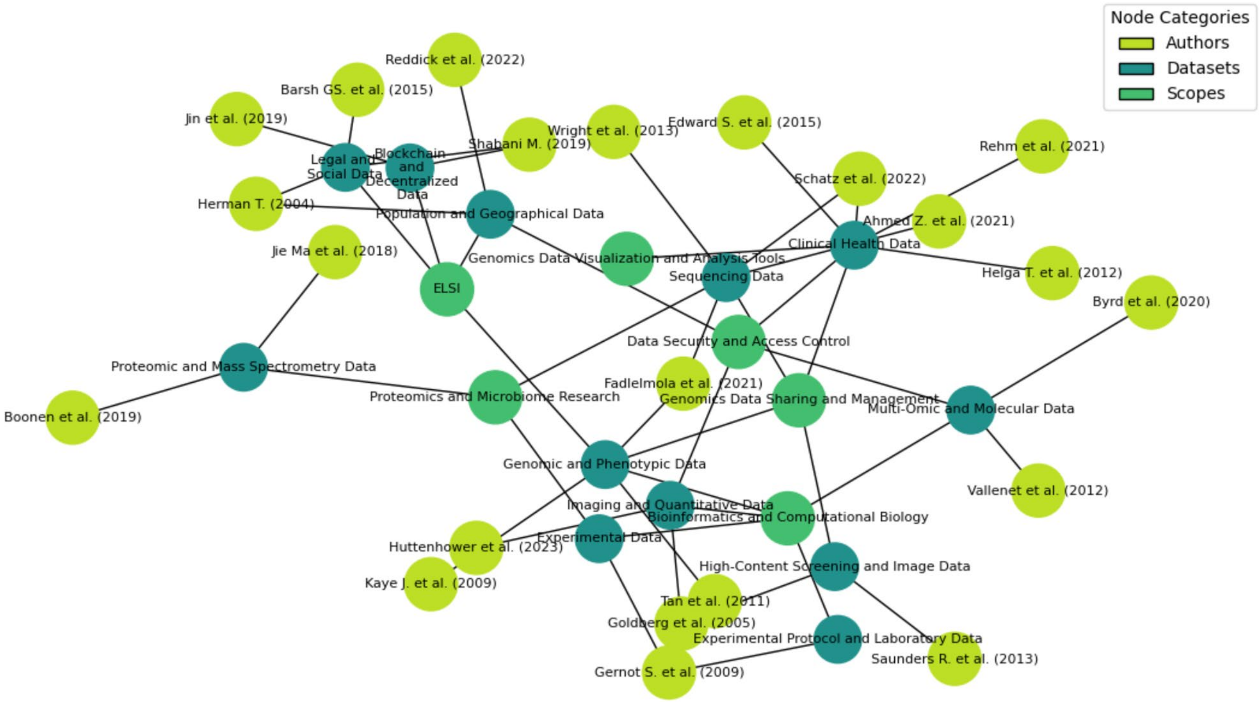


Fig. 4 Network analysis of authors, datasets, and research scopes in ai-based genomics and automated microscopy

of managing heterogeneous data and integrating analysis tools into existing systems. Ethical and privacy considerations emerge prominently, with Wright et al, Reddick et al, and Tavani, discussing the need for culturally sensitive approaches, informed consent, and secure data sharing. Access control and sharing mechanisms are scrutinized by Reddick et al, who emphasize the importance of fine-grained access control and addressing the complexity of sharing large datasets (See Fig. 5).

Data stewardship components: governance measures

Robust data governance procedures are needed in the fields of AI-based genomics and automated microscopy image processing for high-throughput screening research to guarantee data integrity, security, and ethical use. Several authors have contributed insights into key data governance measures, reflecting the diverse range of challenges and considerations innate in managing large-scale datasets in these domains. Schatz et al emphasize standards like GA4GH DUO, versioning, and lineage tracking to maintain data integrity and trace data origins. They also advocate for coding data use terms and involving DACs to promote standardized and responsible data usage practices, fostering collaboration and transparency. Similarly, Fadlelmola et al stress the importance of versioning, lineage tracking, and ethical data-sharing practices. They address ethical challenges in AI-based research, including biases, fairness, and privacy concerns, and propose access control measures to mitigate biases and promote transparency and fairness. Reddick et al advocate for robust security measures and privacy frameworks, such as attribute-based access control, to safeguard data privacy and mitigate bias risks. Wright et al highlight the significance of genomic data governance, including standards compliance, confidentiality measures, and data access agreements (See details in Table 2).

Data quality management practices

The authors examined several aspects of data quality and management across multiple research disciplines, providing insights into how to address relevant difficulties. Schatz et al and Fadlelmola et al emphasize the importance of Data Management Plans (DMPs) in guiding data collection, metadata curation, and data lifecycle management. Stocker, Fischer, et al, and Fadlelmola et al also mentioned the significance of well-curated metadata for understanding and contextualizing heterogeneous data. Furthermore, Wright et al mentioned measures to safeguard sensitive information and reduce the risk of unauthorized access. Reddick et al recommended attribute-based access control models to regulate data sharing

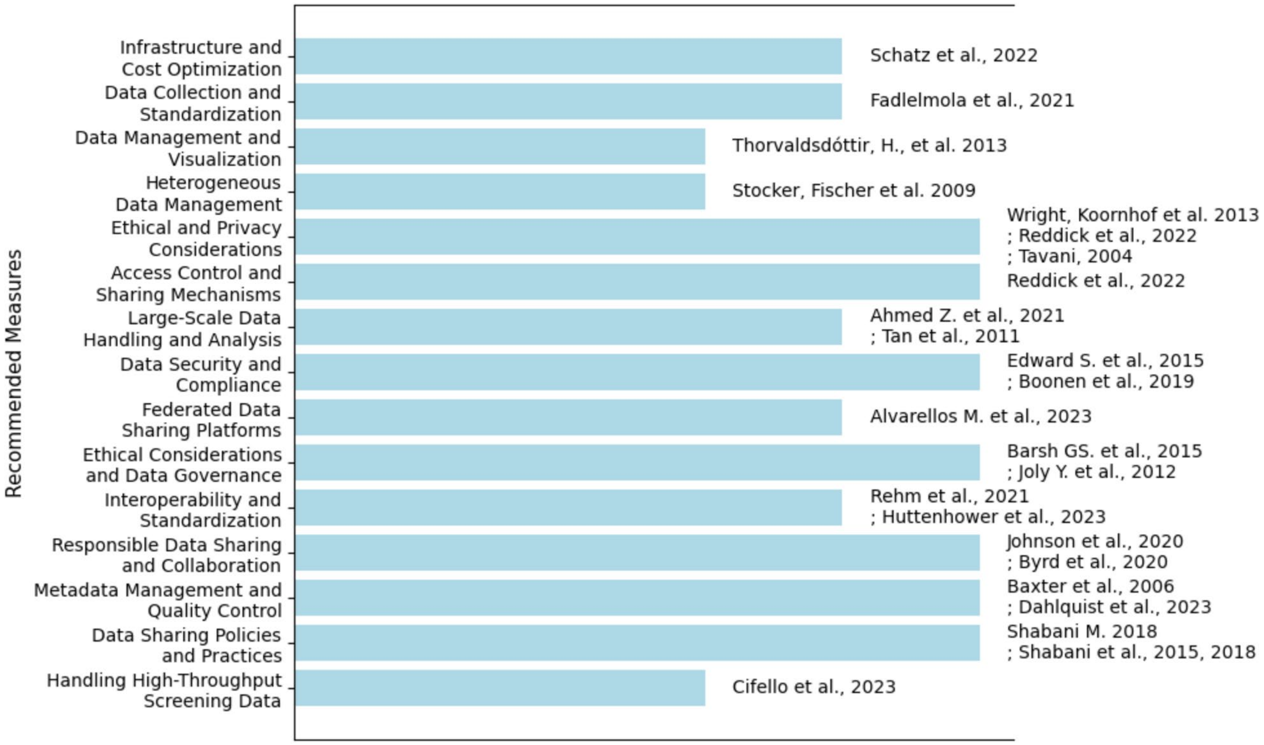


Fig. 5 Key measures reported across various dimensions of data stewardship

Table 2 Data governance measures in AI-based genomics and automated microscopy image analysis for high-throughput screening studies

Theme	Core concepts	Authors contribution
Data integrity and lineage tracking	Maintaining data accuracy and tracing its origin throughout its lifecycle	Schatz et al. (2022) [29] and Fadlelmola et al. (2021) [30] emphasize versioning and lineage tracking to ensure that data can be traced back to its source, preserving reliability and reproducibility
Ethical data use and governance	Addressing ethical considerations, responsible data use, and mitigating biases in AI and genomics research	Fadlelmola et al. (2021) [30] and Rehm et al. (2021) highlight ethical data sharing, considering bias, fairness, and privacy. Reddick et al. (2022) [35] stresses privacy frameworks and ethical guidelines, while Kaye et al. (2009) focus on informed consent and privacy protection in genomic data sharing
Compliance and standards	Adherence to established standards and regulatory frameworks to ensure robust data governance	Schatz et al. (2022) [29], Wright et al. (2013) [31], and Tan et al. (2011) focus on compliance with standards like GA4GH DUO, HIPAA, and other regulatory measures. These studies underscore the importance of standardization for maintaining high data quality and facilitating collaboration
Security and access control	Implementing robust security measures and access controls to protect sensitive data	Reddick et al. (2022) [35] advocate for attribute-based access control to ensure authorized access, reducing data breach risks. Tan et al. (2011) [37] emphasize automated permission systems and restricted access protocols to enhance data security
Privacy and confidentiality	Protecting sensitive data, particularly patient information, from unauthorized access and ensuring confidentiality	Cifello et al. (2023) and Rehm et al. (2021) focus on controlled access to patient data and privacy frameworks to prevent re-identification. Wright et al. (2013) emphasize confidentiality measures within genomic data governance

and ensure fine-grained control over access. Dahlquist et al and Huttenhower et al discuss the need for clear and transparent data governance frameworks to maintain data integrity and transparency. Shabani M explores the potential of blockchain platforms to enhance data security and ownership. Baker et al emphasizes the importance of data curation processes, including compliance with standards, metadata provision, and version control, to maintain data quality. Ma et al highlights the importance of robust infrastructure for efficient data storage, processing, and accessibility. Johnson et al underscores the need to address ethical and legal considerations, ensuring compliance with regulations and safeguarding participant privacy. Finally, Baxter et al address systematic evaluation methods to assess data quality, ensuring accuracy and relevance (See Fig. 6).

Data security and privacy measures

The authors thoroughly investigate different data security and privacy techniques, critical for protecting sensitive information in AI-based genomics and automated microscopy image processing for high-throughput screening studies. Schatz et al explain the importance of robust data security and privacy measures such as data encryption, logging, auditing, intrusion detection, and access controls in reducing risks and safeguarding data authenticity. Other authors, Fadlelmola et al, have discussed data protection rules and responsible data-sharing practices. They argue that data should be restricted, outline data ownership rights, and examine licensing and copyright issues. Authors, Fadlelmola et al reported on data protection rules and responsible data-sharing practices, they argue that data should be restricted, outlining data ownership rights, and examining licensing and copyright issues. Reddick et al advocate for strong data protection and confidentiality requirements, including material transfer agreements, anonymization mechanisms, and data access agreements. Andreas D. et al highlighted the importance of establishing attribute-based access control systems and creating clear guidelines for securely exchanging sensitive data. In Tavani's work, there is a discussion on robust authorization and authentication systems (AAS) to effectively regulate data access, while Edward S. et al address the challenges and risks associated with sharing individual-level genomics data and advocate for stringent security measures to protect privacy. Ahmed Z. et al promote transparent practices and open discussions with cloud service providers to ensure data security and confidentiality in genomic cloud computing environments. Joly Y. et al focus on controlled access mechanisms to balance open access with privacy concerns in genomic data sharing initiatives, while Rehm et al propose a Tiered access system with privacy safeguards for secure and controlled access to sensitive genomic data. Goldberg et al explore advanced cryptographic techniques such as homomorphic cryptography

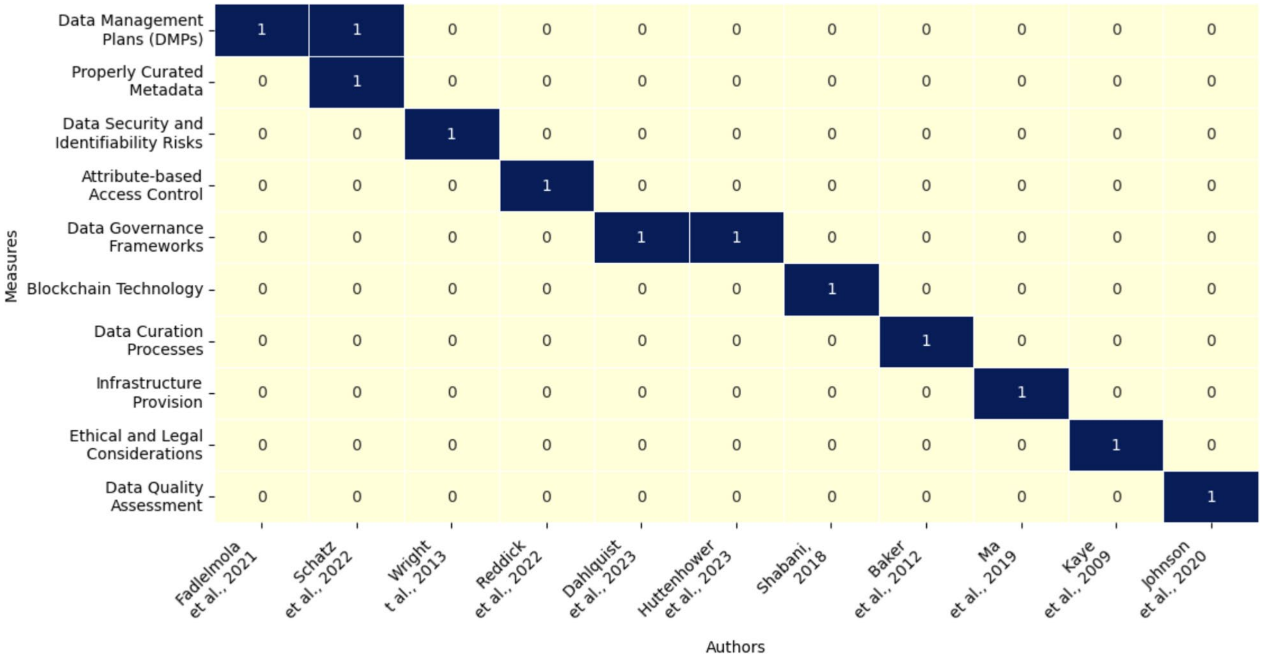


Fig. 6 Heatmap of key measures in data management practices for ai-driven genomics and automated microscopy in high-throughput screening studies

and secure multi-party computation to protect genomic data privacy, and Boonen et al adaptable data management systems and standardized data models in proteomics data sharing initiatives. Lastly, Baker et al highlight stringent security measures to safeguard sensitive genomic and proteomic data, emphasizing robust security protocols and access controls to mitigate security risks and protect sensitive information (See details in Table 3).

Data ownership and access control measures

A significant challenge identified in the literature is achieving a balance between protecting data ownership—whether held by individuals, institutions, or AI models—and ensuring accessibility for research and development. Schatz et al emphasize the importance of role-based access control (RBAC) and data-sharing agreements as mechanisms to facilitate controlled access while safeguarding the ownership rights of data contributors. They further highlight the effectiveness of RBAC, Access Control Lists (ACLs), and encryption as critical tools for managing access control within data governance frameworks. Traditional methodologies such as RBAC are being supplemented by more innovative approaches, including blockchain technology, which offers decentralized and secure mechanisms for data access management.

Table 3 Data privacy and security measures in AI-based genomics and automated microscopy image analysis for high-throughput screening studies

Theme	Core concept	Authors contribution
Robust data security measures	Implementing strong security measures to protect data integrity and prevent unauthorized access	Schatz et al. (2022) emphasize data encryption, logging, auditing, intrusion detection, and access controls to secure data. Baker et al. (2012) also advocate for stringent security measures to safeguard sensitive genomic and proteomic data
Data protection policies and practices	Establishing policies for responsible data sharing, ownership, and licensing	Fadlelmola et al. (2021) highlight the importance of data protection policies, responsible data sharing, and defining data ownership rights. Ahmed Z. et al. (2021) promote transparent practices and open discussions with cloud service providers to ensure data security and confidentiality in genomic cloud computing environments
Confidentiality and privacy safeguards	Ensuring the privacy and confidentiality of sensitive data, particularly in genomics	Reddick et al. (2022) stress confidentiality and advocate for anonymization techniques and material transfer agreements. Edward S. et al. (2015) address the challenges of sharing individual-level genomics data, advocating for stringent security measures to protect privacy. Rehm et al. (2021) propose a tiered access system with privacy safeguards for secure access to sensitive genomic data. Joly Y. et al. (2019) focus on balancing open access with privacy concerns through controlled access mechanisms
Advanced cryptographic techniques	Utilizing cryptographic methods to enhance data security and privacy	Goldberg et al. (2005) explore homomorphic cryptography and secure multi-party computation to protect genomic data privacy. These advanced techniques ensure that data remains secure even during processing and analysis
Authorization and access control	Regulating who can access data through well-defined controls	Tavani, 2004, discusses robust authorization and authentication systems (AAS), while Andreas D. et al. (2014) highlight the implementation of attribute-based access control mechanisms. These measures help in clearly defining user roles and permissions to ensure that only authorized personnel can access sensitive data
Adaptable and standardized data management	Implementing adaptable data management systems and standardized models	Boonen et al. (2019) stress the importance of adaptable data management systems and standardized data models for effective proteomics data sharing. These systems support the dynamic needs of research while ensuring that data remain consistent and secure

Shabani explored the potential of blockchain-based platforms to provide a decentralized approach to access control, offering a flexible and secure alternative to conventional methods. Similarly, Johnson et al advocate for controlled-access models that secure data integrity and facilitate its beneficial utilization in research.

Cifello et al propose a practical framework wherein original data submitters retain ownership, while access is regulated by the National Institutes of Health (NIH), effectively harmonizing ownership with the need for accessibility. In addition to this subtle balance, the literature strongly emphasizes strict adherence to legal and ethical standards, given the sensitive nature of genomic data. Furthermore, compliance with federal regulations, such as the Federal Information Security Management Act (FISMA) and the National Institute of Standards and Technology (NIST) guidelines is crucial to maintaining the integrity and security of data governance frameworks. Material transfer agreements (MTAs) play a pivotal role in ensuring legal compliance during the transfer of genetic samples, thereby safeguarding ownership rights and addressing ethical considerations. Tan et al reported the advantages of automated permission systems, which streamline the access control process and ensure that only authorized entities can engage with sensitive data, reinforcing the integrity of data governance. Finally, the safeguarding of data privacy and security emerges as a key concern, particularly in relation to sensitive genomic data. Various methods, such as encryption, de-identification, and access auditing, are employed to fortify data protection, underscoring the paramount importance of these measures in contexts like AI-based genomics and microscopy image analysis. Schatz et al emphasized the critical role of encryption and data masking techniques in securing data at multiple levels, thereby mitigating the risks associated with unauthorized access. Connor et al discussed the strategic implementation of de-identification measures as a crucial approach to protecting individual privacy while enabling the use of data for research purposes. Byrd et al further underscored the significance of controlled-access repositories in enhancing the security of data-sharing practices, ensuring that sensitive data remains safeguarded throughout the research continuum and that privacy concerns are adequately addressed (See Fig. 7).

Data accountability and traceability measures

Current research on data accountability and traceability in AI-driven genomics and automated microscopy image analysis, particularly within high-throughput screening studies, underscores the imperative for robust governance practices to ensure data integrity, transparency, and compliance. Schatz et al propose an inverted data-sharing model supported by centralized services to improve data ownership and access control, highlighting the necessity for a systematic approach to data governance. Fadlelmola et al emphasize the importance of comprehensive data protection policies and responsible data-sharing practices, particularly in tracking data provenance to maintain transparency and accountability throughout the data lifecycle. Jensen et al stress the need for rigorous accountability and traceability within the Genomic Data Commons (GDC) framework, focusing on controlling

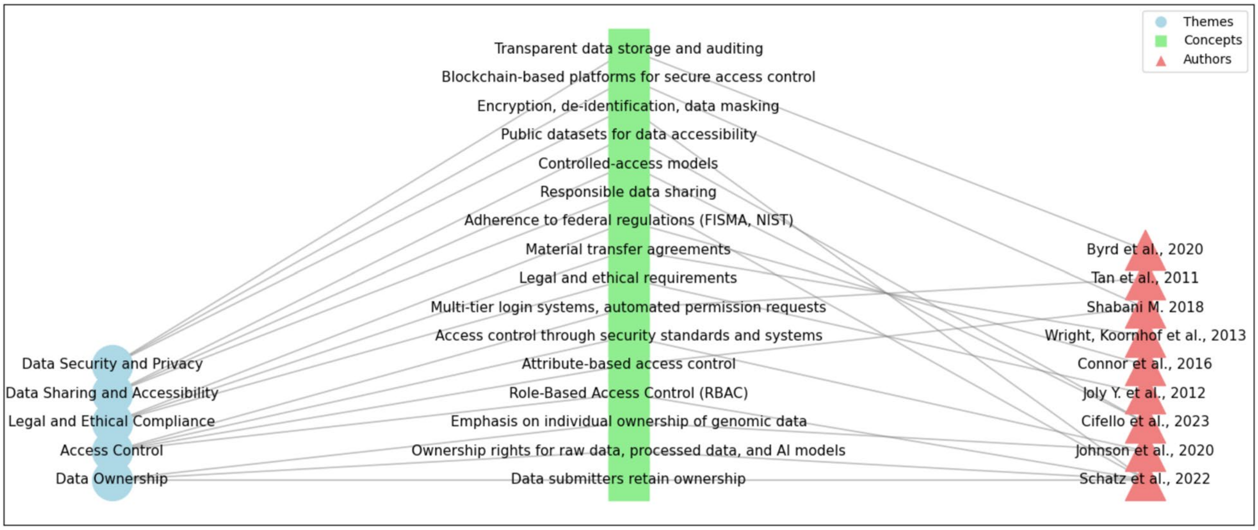

● Themes
■ Concepts
▲ Authors
Transparent data storage and auditing
Blockchain-based platforms for secure access control
Encryption, de-identification, data masking
Public datasets for data accessibility
Controlled-access models
Responsible data sharing
Adherence to federal regulations (FISMA, NIST)
Material transfer agreements
Legal and ethical requirements
Multi-tier login systems, automated permission requests
Access control through security standards and systems
Attribute-based access control
Role-Based Access Control (RBAC)
Emphasis on individual ownership of genomic data
Ownership rights for raw data, processed data, and AI models
Data submitters retain ownership
Data Security and Privacy
Data Sharing and Accessibility
Legal and Ethical Compliance
Access Control
Data Ownership
Byrd et al., 2020
Tan et al., 2011
Shabani M. 2018
Wright, Koornhof et al., 2013
Connor et al., 2016
Joly Y. et al., 2012
Cifello et al., 2023
Johnson et al., 2020
Schatz et al., 2022

Fig. 7 Bipartite network graph of data ownership and access control measures in ai-based genomics and automated microscopy for high-throughput screening studies

access to raw patient sequence data and utilizing the NIH eRA Commons system to enhance traceability.

This perspective aligns with the findings of Jeffrey C. et al, who advocate for meticulous documentation of data management processes, standardized formats, and organized file hierarchies to clarify access control and data ownership. Similarly, Edward S. et al highlight the necessity of transparent practices and thorough documentation to ensure accountability in data ownership and access control mechanisms. Alvarellos M. et al further argue for the establishment of governance bodies to monitor data provenance and lineage, thereby enhancing auditability and ensuring compliance with regulatory standards. Baker et al support this by emphasizing adherence to standards documentation and version control as critical components of data oversight. The complexities surrounding informed consent in data sharing and research collaboration, as discussed by Kaye J. et al, further illustrate the challenges of ensuring data accountability and traceability in research contexts. As a potential solution, Shabani M investigates the application of blockchain technology to enhance transparency and efficiency in data accountability, including the use of smart contracts for compliance and auditability. Jin et al expand on this by highlighting blockchain's role in improving data ownership and access control through immutable tracking of data transactions. Baxter et al underscore the significance of capturing metadata to trace data lineage and provenance, supported by standardized nomenclature and rigorous quality control measures. Johnson et al advocate for data protection policies and controlled-access models, reinforced by clear terms of agreement to ensure authorized access. Finally, Huttenhower et al conclude by emphasizing the necessity for clear and transparent rules governing accountability and traceability, particularly in tracking data provenance and managing versioning to strengthen data governance frameworks (See details in Table 4).

Data curation frameworks and models

Several authors have explored different data curation frameworks that can be used to manage, analyze, and curate genomic and biological data. These frameworks provide researchers with tools for data storage, sharing, and analysis in various research contexts. In a study by Schatz et al the Gen3 AnVIL framework was highlighted as it offers comprehensive data management solutions for genomics research. Fadlelmola et al emphasized the importance of Data Management Plans (DMPs) and related tools in ensuring effective data curation practices. Another study by Thorvaldsdóttir, H., et al focused on the Goby and Integrative Genomics Viewer (IGV) frameworks, which provide functionalities for interactive graphing and data visualization in genomics. Additionally, Stocker, Fischer et al introduced the iLAP framework, specifically designed for laboratory data management, analysis, and protocol development to cater to the unique requirements of experimental settings. Reddick et al proposed an attribute-based access control framework tailored for managing access to genomics data, ensuring data security and privacy. Andreas D. et al discussed SeqBench integrated with an Authorization and Authentication System (AAS) for streamlined data management and access control in sequencing experiments. Tan et al introduced the biology-Related Information Storage Kit (BRISK) framework, focusing on efficient data storage and retrieval in biological research. Jensen et al highlighted the Genomic Data Commons (GDC)

Table 4 Data accountability and traceability measures in AI-based genomics and automated microscopy image analysis for high-throughput screening studies

Theme	Core concept	Authors contribution
Governance models	Strategies and structures for overseeing data sharing and management	Schatz et al. (2022): Inverted data-sharing models, Centralized services. Alvarellos M. et al. (2023): Governance bodies. Jeffrey C. et al. (2023): Documenting data management processes
Technological solutions	Tools and technologies used to enhance data accountability	Shabani M. (2019): Blockchain technology, Smart contracts. Jin et al. (2019): Blockchain. Baxter et al. (2007): Metadata capture
Data provenance and traceability	Methods for documenting and tracking the history of data	Fadlelmola et al. (2021): Tracking data provenance. Cifello et al. (2023): NIH eRA Commons system. Huttenhower et al. (2023): Version control
Compliance and standards	Policies and procedures for adhering to regulatory and quality standards	Edward S. et al. (2015): Transparent practices, documentation. Johnson et al. (2020): Data protection policies. Baker et al. (2012): Standards documentation
Challenges and innovations	Current issues and new approaches in the field	Kaye J. et al. (2009): Informed consent challenges. Dahlquist et al. (2023): Auditing and monitoring

framework, emphasizing centralized data management and controlled access to genomic data resources. Jeffrey C. et al discussed the Harmonization and Integration Pipeline for Functional Genomics (hipFG), offering solutions for harmonizing heterogeneous genomic datasets for integrative analysis. Ahmed Z. et al introduced the Java-based Whole Genome/Exome Sequence Data Processing Pipeline (JWES), offering efficient processing and analysis of high-throughput sequencing data. Alvarellos M. et al emphasized federated data platforms and governance mechanisms for collaborative genomics research initiatives. Schatz et al advocate for the adoption of the GA4GH DRS standard and cloud-agnostic access to facilitate seamless data exchange. Fadlelmola et al emphasize the establishment of formalized genomic data archives and responsible sharing practices within consortia like the H3Africa consortium. Stocker, Fischer et al focus on the development of the iLAP system for lab data management and its integration with repositories and LIMS for enhanced interoperability. Other authors, such as Wright et al, underscore the importance of clear data sharing policies and the establishment of data access committees to ensure responsible data sharing practices. Reddick et al propose an Attribute-Based Access Control (ABAC) model to address challenges in sharing large datasets while maintaining security. Tan et al highlights the development of the BRISK framework for data integration and collaboration in genetics research, including the implementation of automated permissions systems (See Fig. 8).

Interoperability and data sharing practice

Research findings highlight a diverse array of frameworks and initiatives aimed at improving interoperability and data sharing in genomics and biological research. Key strategies include adopting standards such as the GA4GH Data Repository Service (DRS), developing secure access control models like Attribute-Based Access Control (ABAC), and implementing federated data platforms that foster collaborative research.

For instance, Schatz et al describe the Gen3 AnVIL framework, a robust data management platform tailored to genomics. Its integration with the GA4GH DRS standard, support for Dockstore, and cloud-agnostic access significantly enhance interoperability across various cloud platforms, facilitating seamless data sharing and integration. Similarly, Stocker, Fischer et al present the iLAP system, a laboratory information management system (LIMS) designed to streamline data management, analysis, and protocol development. By integrating with various repositories and other LIMS, iLAP boosts interoperability and ensures efficient management and sharing of experimental data. Alvarellos M. et al emphasize the critical role of federated data platforms that align with initiatives like GA4GH and GO FAIR, advocating for the adoption of FAIR (Findable, Accessible, Interoperable, Reusable) principles. These standards enhance secure collaboration among researchers from different institutions.

Furthermore, Reddick et al propose an Attribute-Based Access Control (ABAC) model to manage access to large genomic datasets. This model addresses data security challenges by customizing access controls based on user attributes, ensuring that only authorized individuals can access sensitive genomic information. In addition, Andreas D. et al discuss SeqBench, which integrates an Authorization and Authentication System (AAS) to bolster data management and access control in sequencing experiments. Tan et al introduce the biology-related Information Storage Kit (BRISK) framework, which prioritizes efficient data storage and retrieval in biological research. Its automated permissions systems streamline access control, ensuring that only authorized entities can access sensitive information, a vital aspect in collaborative research environments. In a similar vein, Rehm et al advocate for the adoption of GA4GH standards and cloud-based workflows to enhance data sharing and analysis in genomics research. While cloud solutions improve scalability and accessibility. Lastly, Wright, Koornhof, et al underscore the necessity for clear data-sharing policies and oversight committees to ensure responsible and ethical data sharing, particularly concerning sensitive genomic information. By implementing formalized policies and oversight mechanisms, researchers can ensure that data sharing practices align with legal and ethical standards, while also promoting transparency and accountability (See details in Table 5).

Databases, storage systems, and visualization tools

The research findings indicate that a variety of databases and storage systems are utilized for the management, storage, and dissemination of genomic and biological data. Notably, repositories such as the NCBI Sequence Read Archive (SRA), the EMBL-EBI European Nucleotide Archive, and the DNA Data Bank of Japan (DDBJ) play essential roles in the preservation of raw sequencing data. Concurrently, databases like the NCBI Database of Genotypes and Phenotypes (dbGaP) facilitate the integration and sharing of genomic and phenotypic information. Additionally, systems such as the Integrated Rule-Oriented Data System (iRODS) and the Automated Attribute-Based Access Control (AABAC) model tackle the complexities associated with large-scale data management and access control, thereby ensuring secure and responsible data sharing practices.

In conjunction with these data management systems, visualization tools, including the Integrative Genomics

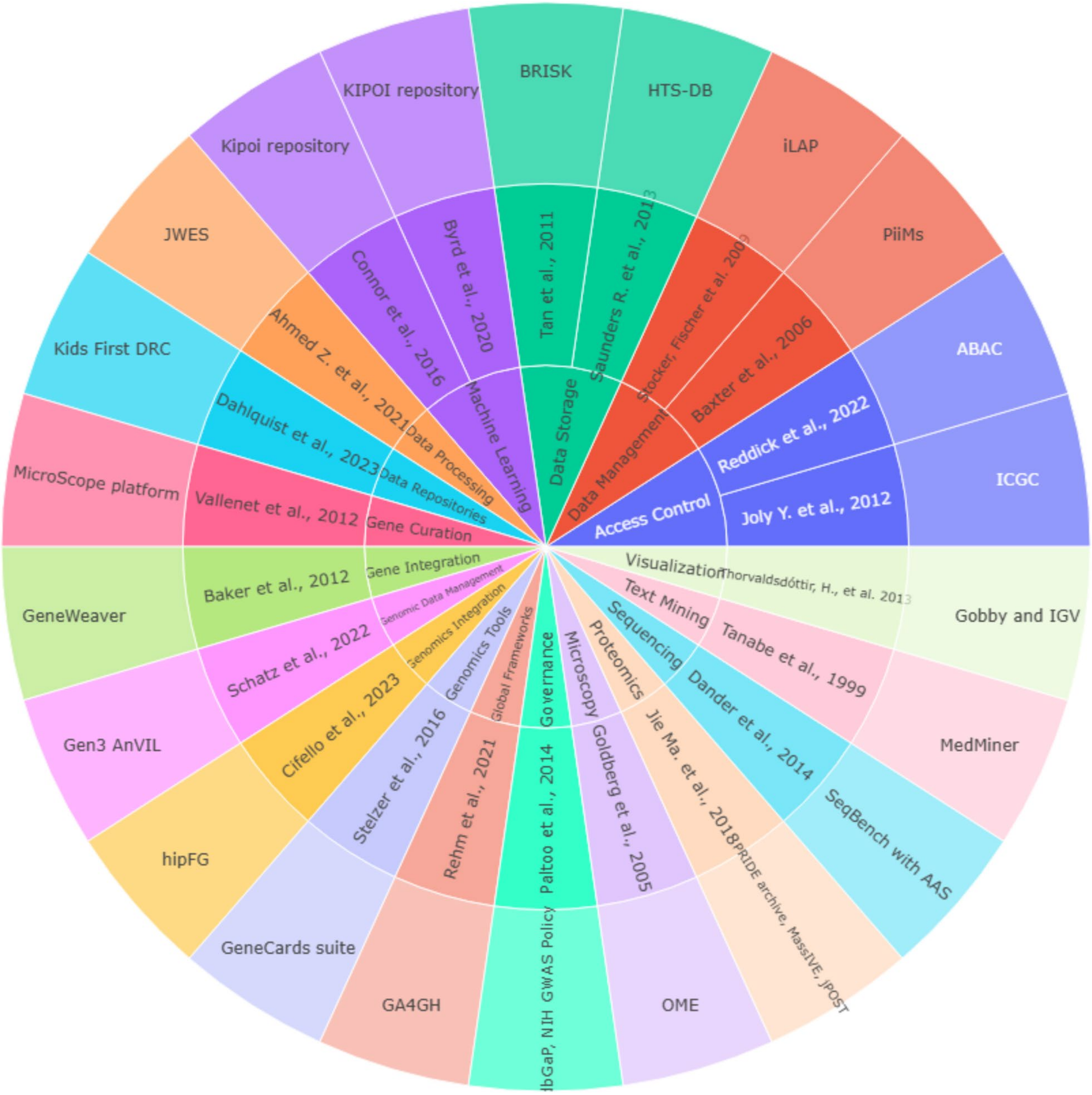


Fig. 8 Sunburst plot of data curation frameworks reported in ai-based genomics and automated microscopy for high-throughput screening studies

Viewer (IGV) and laboratory data management systems like the iLAP system significantly enhance the ability of researchers to analyze and manage data effectively. Specifically, IGV is extensively employed for visualizing and analyzing various genomic data types, integrating metadata with genomic information to support comprehensive data interpretation and exploration. As a result, IGV enhances research insights and supports the discovery of novel genetic relationships. Similarly, the iLAP system, as described by Stocker, Fischer et al, is designed for laboratory data management, analysis, and protocol development. This system seamlessly integrates with repositories and Laboratory Information Management Systems (LIMS), effectively bridging the gap between experimental data generation and storage.

Moreover, specialized platforms such as SeqBench and the Genomic Data Commons (GDC) offer customized solutions for sequencing data management and

Table 5 Interoperability and data sharing measures in AI-based genomics and automated microscopy image analysis for high-throughput screening studies

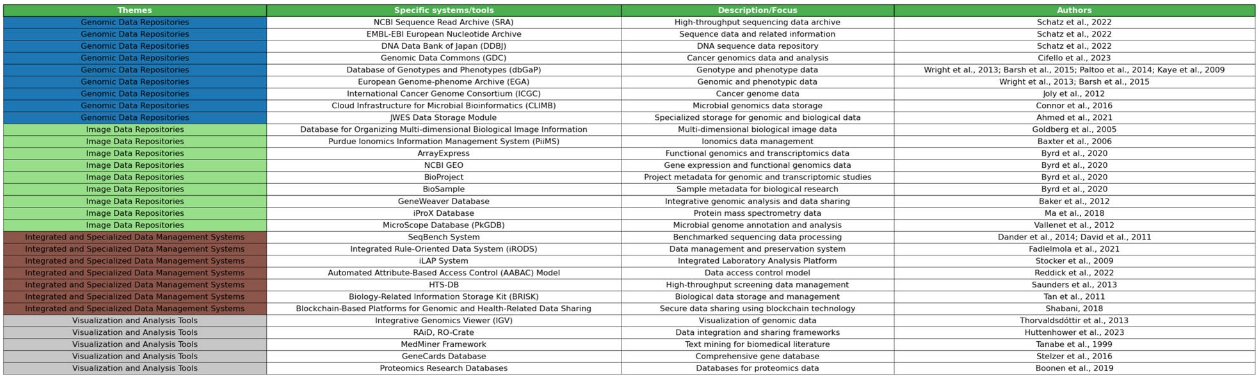
Themes	Core concepts	Authors
Data management frameworks	Adoption of GA4GH DRS standard, Dockstore support, and cloud-agnostic access	Schatz et al., 2022
Data archives and sharing	Establishment of formalized genomic data archives, and responsible sharing practices within consortia like the H3Africa consortium	Fadlelmola et al., 2021
Laboratory data management	Development of the iLAP system for lab data management, integration with repositories, and LIMS for enhanced interoperability	Stocker, Fischer et al., 2009
Data sharing policies	Emphasis on clear data sharing policies and establishment of data access committees	Wright, Koornhof et al., 2013
Access control models	Proposal of Attribute-Based Access Control (ABAC) model for secure large dataset sharing	Reddick et al., 2022
Data integration frameworks	Development of the BRISK framework for data integration and collaboration in genetics research, including automated permissions systems	Tan et al., 2011
Federated data platforms	Promotion of federated data platform interoperability, and support for GA4GH and GO FAIR initiatives	Alvarellos M. et al., 2023
Standards adoption	Advocacy for GA4GH standards adoption and cloud-based workflows for data sharing and analysis	Rehm et al., 2021
Standardized data outputs	Utilization of platforms like GeneWeaver and Ontological Discovery Environment to ensure standardized data outputs and promote interoperability	Baker et al., 2012

controlled data access. For instance, SeqBench, introduced by Dander et al and David W. et al, provides tools for processing and analyzing high-throughput sequencing data, which is vital for advancing research in genomics and personalized medicine. In parallel, Cifello et al highlight the GDC's role as a centralized platform for genomic data management, supporting controlled access to genomic data resources (See Fig. 9).

Discussion

The review consolidated findings related to data stewardship and curation within the context of AI-driven technology and research, targeting researchers, policymakers, and relevant stakeholders. Its objective was to elucidate the theoretical foundations, practical implications, and obstacles associated with data management in the fields of genomics, proteomics, microbiome research, and AI technologies. Among the principal elements highlighted were data governance, quality assurance, privacy and security protocols, ownership rights, access management, accountability, traceability, curation methodologies, and database/storage infrastructures.

The results of this review highlight the diverse array of data types essential for AI-driven genomics and microscopy image analysis in high-throughput screening research. These data types, which encompass sequencing data, high-content screening, and image data, are foundational to advancements in the field. For instance, genomic data, such as single-cell sequencing data, provides insights into cellular complexity and heterogeneity, facilitating precise mapping of cell types and states [10]. Similarly, proteomic data derived from deep visual proteomics techniques allows for an unbiased characterization of cellular functions and the identification



Themes	Specific systems/tools	Description/Focus	Authors
Genomic Data Repositories	NCBI Sequence Read Archive (SRA)	High-throughput sequencing data archive	Schatz et al., 2022
Genomic Data Repositories	EMBL-EBI European Nucleotide Archive	Sequence data and related information	Schatz et al., 2022
Genomic Data Repositories	DNA Data Bank of Japan (DDBJ)	DNA sequence data repository	Schatz et al., 2022
Genomic Data Repositories	Genomic Data Commons (GDC)	Cancer genomics data and analysis	Cifello et al., 2023
Genomic Data Repositories	Database of Genotypes and Phenotypes (dbGaP)	Genotype and phenotype data	Wright et al., 2013; Barsh et al., 2015; Paltoo et al., 2014; Kaye et al., 2009
Genomic Data Repositories	European Genome-phenome Archive (EGA)	Genomic and phenotypic data	Wright et al., 2013; Barsh et al., 2015
Genomic Data Repositories	International Cancer Genome Consortium (ICGC)	Cancer genome data	Joly et al., 2012
Genomic Data Repositories	Cloud Infrastructure for Microbial Bioinformatics (CLIMB)	Microbial genomics data storage	Connor et al., 2016
Genomic Data Repositories	JWES Data Storage Module	Specialized storage for genomic and biological data	Ahmed et al., 2021
Image Data Repositories	Database for Organizing Multi-dimensional Biological Image Information	Multi-dimensional biological image data	Goldberg et al., 2005
Image Data Repositories	Purdue Ionomics Information Management System (PiiMS)	Ionomics data management	Baxter et al., 2006
Image Data Repositories	ArrayExpress	Functional genomics and transcriptomics data	Byrd et al., 2020
Image Data Repositories	NCBI GEO	Gene expression and functional genomics data	Byrd et al., 2020
Image Data Repositories	BioProject	Project metadata for genomic and transcriptomic studies	Byrd et al., 2020
Image Data Repositories	BioSample	Sample metadata for biological research	Byrd et al., 2020
Image Data Repositories	GeneWeaver Database	Integrative genomic analysis and data sharing	Baker et al., 2012
Image Data Repositories	iProX Database	Protein mass spectrometry data	Ma et al., 2018
Image Data Repositories	MicroScope Database (PkGDB)	Microbial genome annotation and analysis	Vallenet et al., 2012
Integrated and Specialized Data Management Systems	SeqBench System	Benchmarked sequencing data processing	Dander et al., 2014; David et al., 2011
Integrated and Specialized Data Management Systems	Integrated Rule-Oriented Data System (iRODS)	Data management and preservation system	Fadlelmola et al., 2021
Integrated and Specialized Data Management Systems	iLAP System	Integrated Laboratory Analysis Platform	Stocker et al., 2009
Integrated and Specialized Data Management Systems	Automated Attribute-Based Access Control (AABAC) Model	Data access control model	Reddick et al., 2022
Integrated and Specialized Data Management Systems	HTS-DB	High-throughput screening data management	Saunders et al., 2013
Integrated and Specialized Data Management Systems	Biology-Related Information Storage Kit (BRISK)	Biological data storage and management	Tan et al., 2011
Integrated and Specialized Data Management Systems	Blockchain-Based Platforms for Genomic and Health-Related Data Sharing	Secure data sharing using blockchain technology	Shabani, 2018
Visualization and Analysis Tools	Integrative Genomics Viewer (IGV)	Visualization of genomic data	Thorvaldsdóttir et al., 2013
Visualization and Analysis Tools	RAiD, RO-Crate	Data integration and sharing frameworks	Huttenhower et al., 2023
Visualization and Analysis Tools	MedMiner Framework	Text mining for biomedical literature	Tanabe et al., 1999
Visualization and Analysis Tools	GeneCards Database	Comprehensive gene database	Stelzer et al., 2016
Visualization and Analysis Tools	Proteomics Research Databases	Databases for proteomics data	Boonen et al., 2019

Fig. 9 overview of databases, storage systems, and visualization tools in ai-based genomics and automated microscopy for high-throughput screening studies

of disease-associated protein markers [11]. Similarly, proteomic data derived from deep visual proteomics techniques allows for an unbiased characterization of cellular functions and the identification of disease-associated protein markers [12]. Furthermore, microbiome data, which includes profiles associated with systemic and tumor microenvironments, plays a significant role in influencing cancer development and treatment responses, with AI-driven computational pathology systems offering valuable insights for clinical decision-making [21]. The integration of these varied data types enhances our understanding of cellular functions, disease mechanisms, and treatment strategies, underscoring the critical role of data-driven methodologies in high-throughput screening research.

The challenges identified in this review closely align with those noted in existing literature within the field [38, 39]. Various authors have pointed out common obstacles in data stewardship, such as effective data management, ensuring data quality, addressing privacy and security concerns, and tackling issues of bias and fairness. Additionally, challenges related to the interpretability and explainability of AI systems, as well as the acquisition of technical expertise, have emerged as recurring themes. The necessity of integrating, comparing, and visualizing large, multi-dimensional datasets from diverse sources has been emphasized as vital for unlocking the potential of AI across various research domains [40]. Recommendations for addressing these challenges include strategies focused on data quality, volume, privacy, security, bias, interpretability, explainability, and technical expertise [41]. Moreover, EU researchers face specific challenges in reconciling data protection requirements and AI research, particularly concerning the processing of large-scale databases containing personal data. Responsible data integration in machine learning pipelines necessitates concerted efforts to address concerns about data quality and bias, leading to the development of techniques and methods that optimize the principles of responsible data science in data integration tasks. Overall, the convergence of findings from this review and existing literature emphasizes the multifaceted nature of challenges in data stewardship within AI-driven research contexts, highlighting the need for comprehensive and nuanced approaches to address these complexities.

The review identified a comprehensive set of data privacy and security measures employed in AI-driven genomics and microscopy image analysis, which align with strategies proposed in the literature. These measures include data encryption, responsible data sharing practices, and authentication systems with defined user roles were commonly observed. Authors also emphasized the importance of open discussions with cloud service providers and compliance with industry-recognized data protection frameworks to ensure data security and confidentiality, particularly in genomic cloud computing.

Furthermore, the review highlighted the significance of federated learning strategies and adherence to data governance principles as essential for balancing open access with privacy concerns. Advanced cryptographic techniques, such as homomorphic cryptography and secure multi-party computation, were proposed as innovative solutions to safeguard sensitive genomic data. In the field of genomics, mechanisms like differential privacy were suggested as a means to share aggregated statistical information while preserving privacy, thereby addressing vulnerabilities to inference attacks [42, 43]. Similarly, in the domain of microscopy image analysis, federated learning algorithms, such as FedTransfer, were introduced to enhance model generalization while ensuring privacy and security [44]. Additionally, data sharing strategies based on style transfer were proposed to mitigate performance penalties caused by data distribution differences among users [45, 46]. These approaches collectively aim to provide privacy guarantees, protect individual participants, and facilitate collaborative research in a distributed and secure manner.

The review findings regarding data ownership and access control measures resonate with existing literature, highlighting the importance of robust frameworks for safeguarding sensitive genomic data. Commonly reported measures, such as user authentication, role-based access control (RBAC), and encryption, are consistent with strategies discussed in prior studies. For instance, Reddick et al [35] emphasized the significance of RBAC in controlling access to genomic data, particularly in research settings where multiple users may require varying levels of access.

Similarly, the use of data sharing agreements, access control lists (ACLs), and data access auditing reflects a collective emphasis on regulating data access and ensuring compliance with federal regulations. Rosa et al [47] explored the role of data sharing agreements in delineating responsibilities and permissions among collaborators, ensuring transparent data sharing practices. Meanwhile, Dyke [48] highlighted the importance of comprehensive data access auditing mechanisms to monitor and track data access activities, enabling accountability and compliance with regulatory requirements. However, while our review focuses on automated attribute-based access control (AABAC) models and extensions of the XACML framework, other studies may have explored alternative approaches or emphasized different aspects of access control. Federated data platforms and blockchain technology have been explored to enhance data security and traceability in genomic research settings. Federated data platforms, such as those discussed by Alvarellos et al. [49] and Dervishi et al. [50] enable secure data sharing without physically moving the data, allowing for global collaboration and representation of diverse populations. On the other hand, blockchain-based frameworks like PGxChain, proposed by Albalwy et al. utilize smart contracts to ensure secure and equitable access to genomic data, addressing privacy concerns and enabling interoperability between multiple healthcare providers [51]. Additionally, Visscher et al. [52] and Manzoor et al. [53] propose decentralized and privacy-preserving systems that combine homomorphic encryption, zero-knowledge proofs, and blockchain to protect sensitive genetic data and enable verifiability. These approaches offer potential solutions to the challenges of data security and privacy in genomic research, facilitating data sharing and analysis while maintaining confidentiality and trust. Furthermore, variations may arise in the implementation of data access policy models, such as open access, controlled access, and registered access, depending on institutional requirements and regulatory frameworks.

Conclusions

This study highlights the progress and ongoing challenges in data stewardship within AI-driven genomics and automated microscopy image analysis. Key advancements include the adoption of standards like GA4GH DUO and effective versioning practices, which have improved data integrity and lineage tracking. Ethical data-sharing practices—focused on addressing bias, fairness, and privacy—are supported by technologies such as role-based access control (RBAC) and blockchain. However, challenges persist, particularly in integrating diverse data sources, ensuring interoperability across platforms, and maintaining high data quality. Balancing data ownership with the need for accessible research, especially in genomics, remains a significant issue, further complicated by complex legal and ethical considerations.

Data curation frameworks like Gen3 AnVIL, iLAP, and advanced metadata techniques are essential for managing genomic data, but interoperability and data quality issues still pose hurdles. In terms of security, cryptographic methods like homomorphic encryption and federated learning protect privacy in collaborative research, though concerns over unauthorized access and data breaches persist. The use of attribute-based access control (ABAC) and automated permission systems streamlines collaboration, yet complexities in data-sharing agreements and material transfer agreements (MTAs) remain challenging. Adherence to regulations like HIPAA is crucial for maintaining both data accessibility and privacy. Blockchain enhances data accountability and traceability, but ensuring transparency and ethical compliance, particularly in areas like informed consent and data provenance, continues to require careful management.

Limitations and implications of the study

Although this review offers important insights into data stewardship and curation practices in AI-driven genomics and microscopy, there are several limitations. The search was limited to articles published up to January 2024, potentially missing the latest developments. Additionally, the focus on English-language studies and the exclusion of grey literature may introduce biases. The reliance on database searches may also have overlooked relevant research from other sources. Finally, despite efforts to ensure rigor, the subjective nature of data extraction and synthesis may introduce some bias into the findings.

Despite these limitations, the findings of this study have important implications for both researchers and policymakers. First, it underlines the need for continuous innovation in data stewardship, particularly in developing more advanced curation tools and access control mechanisms to manage the huge nature of data. Second, the study points out that while existing legal and ethical frameworks provide a foundation for data governance, these frameworks must evolve to address new challenges introduced by AI technologies. Policymakers and researchers must collaborate to create governance models that are both effective and adaptable to the rapidly changing landscape of AI-driven research. Lastly, the study underscores the importance of fostering transparency, accountability, and ethical responsibility in research practices to build trust in AI applications. This will be critical for advancing ethically sound and scientifically robust innovations in genomics and microscopy.

Abbreviations

AABAC	Automated attribute-based access control
AAS	Authorization and authentication system
ABAC	Attribute-based access control
AI	Artificial intelligence
BRISK	Biology-related information storage kit
CASPE	Critical appraisal skills programme for EBM
DDBJ	DNA Data Bank of Japan
dbGaP	Database of genotypes and phenotypes
DMPs	Data management plans
DUO	Data use ontology
FAIR	Findable, accessible, interoperable, reusable
GA4GH	Global alliance for genomics and health
GDPR	General data protection regulation
GDC	Genomic data commons
HCS	High-content screening
hipFG	Harmonization and integration pipeline for functional genomics
HTS	High-throughput screening
HTS-DB	High-throughput screening database
IGV	Integrative genomics viewer
iLAP	Laboratory data management, analysis, and protocol development
iRODS	Integrated rule-oriented data system

JWES	Java-based whole genome/exome sequence data processing pipeline
MIAME	Minimum information about a microarray experiment
MIBI	Minimum information in biological imaging
NFT	Non-fungible token
NGS	Next-generation sequencing
OME	Open microscopy environment
RBAC	Role-based access control
SeqBench with AAS	SeqBench with authorization and authentication system
SRA	Sequence read archive

Author contribution

Asefa Adimasu Taddese and Assefa Chekole contributed equally to the conception and design of the study, literature review, data extraction, thematic analysis, and synthesis of results. Asefa Adimasu Taddese led the writing of the manuscript, with critical input and revisions provided by Assefa Chekole. Both authors approved the final version of the manuscript for submission.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Competing interest

The authors declare no competing interests.

Author details

[1]Academy of Wellness and Human Development, Faculty of Arts and Social Sciences, Hong Kong Baptist University, Hong Kong SAR, China. [2]Department of Information Science, College of Informatics, University of Gondar, Gondar, Ethiopia. [3]Dr. Stephen Hui Research Centre for Physical Recreation and Wellness, Faculty of Arts and Social Sciences, Hong Kong Baptist University, Hong Kong SAR, China.

Received: 4 April 2024 Accepted: 9 January 2025
Published online: 23 February 2025

References

1. Wright GE, Koornhof PG, Adeyemo AA, Tiffin NJBME. Ethical and legal implications of whole genome and whole exome sequencing in African populations. BMC Medical Ethics. 2013;14:1–15.
2. Schilling MP, El Khaled R, Faraj E, Gómez JEU, Sonnentag SJ, Wang F, Nestler B, et al. Automated high-throughput image processing as part of the screening platform for personalized oncology. Scient Rep. 2023. https://doi.org/10.1038/s41598-023-32144-z.
3. Toczydlowski RH, Liggins L, Gaither MR, Anderson TJ, Barton RL, Berg JT, et al. Poor data stewardship will hinder global genetic diversity surveillance. Proceedings of the National Academy of Sciences. 2021;118(34): e2107934118.
4. Wildenhain J, editor Application of multivariate statistics and machine learning to phenotypic imaging and chemical high-content data2016.
5. Bock C, Datlinger P, Chardon F, Coelho MA, Dong MB, Lawson KA, Tian L, Maroc L, Norman TM, Song B, Stanley G, Chen S, Garnett M, Li W, Moffat J, Qi LS, Shapiro RS, Shendure J, Weissman JS, Zhuang X. High-content CRISPR screening. Nature Reviews Methods Primers. 2022. https://doi.org/10.1038/s43586-021-00093-4.
6. Guo P, Chen LP, Chen W. Advances in high-content screening applications in toxicology research. Zhonghua yu fang yi xue za zhi [Chinese journal of preventive medicine]. 2022;56(1):15–9.
7. Huang L, Zhang X, Feng Y, Liang F, Wang W. High content drug screening of primary cardiomyocytes based on microfluidics and real-time ultra-large-scale high-resolution imaging. Lab Chip. 2022;22(6):1206–13.
8. Faenza A, Bocchi M, Pecorari N, Franchi E, Guerrieri R. Impedance measurement technique for high-sensitivity cell detection in micro-structures with non-uniform conductivity distribution. Lab on a Chip. 2012;12(11):2046–52.
9. Love MS, McNamara CW. High-content screening for cryptosporidium drug discovery. Methods in molecular biology (Clifton, NJ). 2020;2052:303–17.
10. André O, Kumra Ahnlide J, Norlin N, Swaminathan V, Nordenfelt P. Data-driven microscopy allows for automated context-specific acquisition of high-fidelity image data. Cell Reports Methods. 2023;3(3): 100419.
11. Sun T, Niu X, He Q, Chen F, Qi RQ. Artificial Intelligence in microbiomes analysis: A review of applications in dermatology. Front Microbiol. 2023;14:1112010.
12. Steigele S, Siegismund D, Fassler M, Kustec M, Kappler B, Hasaka T, et al. Deep learning-based HCS image analysis for the enterprise. SLAS discovery : advancing life sciences R & D. 2020;25(7):812–21.
13. Zhang W, Suo J, Yan Y, Yang R, Lu Y, Jin Y, et al. iSMOD: an integrative browser for image-based single-cell multi-omics data. Nucleic Acids Res. 2023;51(16):8348–66.
14. Giansanti V, Giannese F, Botrugno OA, Gandolfi G, Balestrieri C, Antoniotti M, et al. Scalable integration of multiomic single cell data using generative adversarial networks. Bioinformatics. 2023;39:825.
15. Maitra C, Seal DB, Das V, De RK. Unsupervised neural network for single cell Multi-omics INTegration (UMINT): an application to health and disease. Front Mol Biosci. 2023;10:1184748.
16. Lu P, Oetjen KA, Oh ST, Thorek DLJJb. Interpretable spatial cell learning enhances the characterization of patient tissue microenvironments with highly multiplexed imaging data. Biorxiv. 2023;19:284.
17. Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. Genome Medicine. 2019;11(1):70.
18. Singh S, Carpenter AE, Genovesio A. Increasing the content of high-content screening: an overview. Journal of Biomolecular Screening. 2014;19(5):640–50.
19. Chai B, Efstathiou C, Yue H, Draviam VM. Opportunities and challenges for deep learning in cell dynamics research. Trends in Cell Biology. 2024;34(11):955–67. https://doi.org/10.1016/j.tcb.2023.10.010.
20. Kemmer I, Keppler A, Serrano-Solano B, Rybina A, Özdemir B, Bischof J, et al. Building a FAIR image data ecosystem for microscopy communities. Histochem Cell Biol. 2023;160(3):199–209.
21. Mund A, Coscia F, Kriston A, Hollandi R, Kovács F, Brunner AD, et al. Deep Visual Proteomics defines single-cell identity and heterogeneity. Nat Biotechnol. 2022;40(8):1231–40.
22. Ahmed M, Kim DR. Editorial: Opportunities and challenges in reusing public genomics data. Front Pharmacol. 2023;14:1226756.
23. Inau ET, Sack J, Waltemath D, Zeleke AA. Initiatives, concepts, and implementation practices of the findable, accessible, interoperable, and reusable data principles in health data stewardship: scoping review. J Med Internet Res. 2023;25: e45013.
24. Johns M, Meurers T, Wirth FN, Haber AC, Müller A, Halilovic M, et al. Data provenance in biomedical research: scoping review. J Med Internet Res. 2023;25: e42289.
25. Eismann B, Krieger TG, Beneke J, Bulkescher R, Adam L, Erfle H, Herrmann C, Eils R, Conrad C. Automated 3D light-sheet screening with high spatiotemporal resolution reveals mitotic phenotypes. Journal of Cell Science. 2020. https://doi.org/10.1242/jcs.245043.
26. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet. 2001;29(4):365–71.
27. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3(1): 160018.
28. Pawson R, Greenhalgh T, Harvey G, Walshe K. Realist review – a new method of systematic review designed for complex policy interventions. J Health Serv Res Policy. 2005;10:21–34.
29. Schatz MC, Philippakis AA, Afgan E, Banks E, Carey VJ, Carroll RJ, et al. Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space. Cell Genomics. 2022;2(1):100085. https://doi.org/10.1016/j.xgen.2021.100085.
30. Fadlelmola FM, Zass L, Chaouch M, Samtal C, Ras V, Kumuthini J, et al. Data Management Plans in the genomics research revolution of Africa: Challenges and recommendations. J Biomed Inform. 2021;122: 103900.
31. Wright GE, Koornhof PG, Adeyemo AA, Tiffin N. Ethical and legal implications of whole genome and whole exome sequencing in African populations. BMC Med Ethics. 2013;14:21.

32. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14(2):178–92.
33. Saunders RE, Instrell R, Rispoli R, Jiang M, Howell M. HTS-DB: an online resource to publish and query data from functional genomics high-throughput siRNA screening projects. Database. 2013. https://doi.org/10.1093/database/bat072.
34. Jin XL, Zhang M, Zhou Z, Xiaoyu Yu. Application of a blockchain platform to manage and secure personal genomic data: a case study of lifeCODE.ai in China. Journal of Medical Internet Research. 2019;21(9):e13587. https://doi.org/10.2196/13587.
35. Reddick D, Presley J, Feltus FA, Shannigrahi S. WiP: AABAC - Automated Attribute Based Access Control for Genomics Data. Proceedings of the 27th ACM on Symposium on Access Control Models and Technologies; New York, NY, USA: Association for Computing Machinery; 2022. p. 217–22.
36. Spahic D, Mauša G, Pavelić SK, Galinac Grbac TJtICoI, Communication Technology E, Microelectronics. Data Storage and Analysis system for conducting Biotechnological Experiments. 2015:470–5.
37. Tan A, Tripp B, Daley D. BRISK–research-oriented storage kit for biology-related data. Bioinformatics (Oxford, England). 2011;27(17):2422–5.
38. Khosravi H, Sadiq S, Amer-Yahia S. Data management of AI-powered education technologies: Challenges and opportunities. Learning Letters. 2023;1:2.
39. Williamson HF, Brettschneider J, Caccamo M, Davey RP, Goble C, Kersey PJ, May S, Morris RJ, Ostler R, Pridmore T, Rawlings C, Studholme D, Tsaftaris SA, Leonelli S. Data management challenges for artificial intelligence in plant and agricultural research. F1000Research. 2021;10:324. https://doi.org/10.12688/f1000research.52204.1.
40. Nargesian F, Asudeh A, Jagadish HVJPotSAICoWS, Mining D. Next-generation Challenges of Responsible Data Integration. 2023.
41. Aldoseri A, Al-Khalifa KN, Hamouda AM. Re-thinking data strategy and integration for artificial intelligence: concepts, opportunities, and challenges. Applied Sciences. 2023;13(12):7082.
42. Cheng C, Liu Z, Zhao F, Wang X, Wu F, editors. Security Protection of Research Sensitive Data Based on Blockchain. 2022 21st International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES); 2022 14–18 Oct. 2022.
43. Zixiang L, Cheng C, Feng Z, Xiang W, Feng WJsISoDC, Engineering AfB, et al. Application of Ship Data Based on Blockchain. 2022:229–32.
44. Alserr NA, Kale G, Mutlu O, Tastan O, Ayday EJPotA-Pbc. Tuning Privacy-Utility Tradeoff in Genomic Studies Using Selective SNP Hiding. 2023;2023.
45. Torkzadehmahani R, Nasirigerdeh R, Blumenthal DB, Kacprowski T, List M, Matschinske J, Spaeth J, Wenke NK, Baumbach Jan. Privacy-preserving artificial intelligence techniques in biomedicine. Methods of Information in Medicine. 2022;61(S 01):e12–27. https://doi.org/10.1055/s-0041-1740630.
46. Ma X, Yang RW-M, Zheng MJtICoM, Sensing, Networking. RDP-WGAN: Image Data Privacy Protection Based on Rényi Differential Privacy. 2022:320–4.
47. Rosa M, Cerbo FD, Lozoya RCJPottASoACM, Technologies. Declarative Access Control for Aggregations of Multiple Ownership Data. 2020.
48. Dyke SOM, editor Genomic data access policy models2020.
49. Alvarellos M, Sheppard HE, Knarston I, Davison C, Raine N, Seeger T, et al. Democratizing clinical-genomic data: How federated platforms can promote benefits sharing in genomics. Front Genet. 2022;13:1045450.
50. Albalwy F, McDermott JH, Newman WG, Brass A, Davies A. A blockchain-based framework to support pharmacogenetic data sharing. Pharmacogenomics J. 2022;22(5–6):264–75.
51. Visscher L, Alghazwi M, Karastoyanova D, Turkmen FJPotASCoC, Security C. Poster: Privacy-preserving Genome Analysis using Verifiable Off-Chain Computation. 2022.
52. Dervishi L, Wang X, Li W, Halimi A, Vaidya J, Jiang X, et al. Facilitating Federated Genomic Data Analysis by Identifying Record Correlations while Ensuring Privacy. AMIA Annual Symposium proceedings AMIA Symposium. 2022;2022:395–404.
53. Manzoor S, Gouglidis A, Bradbury M, Suri NJPotASCoC, Security C. Poster. 2022.

Publisher's Note