Discover Artificial Intelligence

REVIEW Open Access

Synthetic data generation in genomic cancer medicine: a review of global research trends in the last ten years

Valentina De Nicoló[1,2*], Maria Frasca[1], Agnese Graziosi[1], Gianluca Gazzaniga[1], Davide La Torre[3] and Arianna Pani[1]

*Correspondence:
Valentina De Nicoló
valentina.denicolo@unimi.it
[1]Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy
[2]Department of Public Health and Infectious Diseases, University of Rome, Rome, Italy
[3]SKEMA Business School, Université Côte d'Azur, Sophia Antipolis, France

Abstract
This paper reviews the global evolution of synthetic data (SD) generation in the field of genomic cancer medicine, with an analysis of research trends from the past decade. The use of artificial intelligence, particularly machine learning and deep learning techniques has transformed this area, providing solutions to overcome the limited availability of real clinical data. Through a bibliometric analysis of a wide sample of scientific articles from SCOPUS, this study highlights the adoption of SD generation techniques in oncological applications, focusing on major methodologies and challenges. Key application areas, such as multi-omics integration (genomics, transcriptomics, and proteomics) and tumor genomic heterogeneity, emerge as fields of growing interest. Despite noise management and performance optimization challenges, advanced machine learning techniques prove essential for generating high-quality SD that reflects biological complexity. The study also identifies key open challenges, such as simulation accuracy and noise control, offering insights into future applications of SD in personalized medicine and cancer therapy.

Keywords Synthetic data, Genomic medicine, Cancer research, Data privacy, Machine learning

1 Introduction

The incorporation of genomic insights into the field of cancer medicine has precipitated a paradigm shift in our understanding and approach to combating this disease. By analysing the genetic composition of tumors, researchers can identify biomarkers that predict disease progression, therapeutic responses and patient outcomes, thereby enhancing the precision and personalisation of cancer care [25]. Nevertheless, the considerable potential of genomic data also gives rise to intricate challenges on privacy and data accessibility. Genomic data are inherently unique to individuals, which subjects it to rigorous privacy regulations that limit data sharing and often impede the collaborative efforts needed to advance precision oncology [10].

The generation of SD has recently emerged as an innovative solution to bridge this gap in genomic cancer research. Despite the availability of large-scale genomic repositories,

SD generation remains essential for several reasons. First, while large-scale genomic repositories provides an extensive dataset, it is still limited by patient selection biases, underrepresentation of certain ethnic groups, and the lack of real-time data updates. Moreover, large-scale genomic repositories are often subject to strict data usage policies, limiting accessibility for large-scale AI-driven studies. In contrast to conventional anonymisation techniques, which may still leave data vulnerable to re-identification, the generation of SD offers a fundamentally different approach [16]. This involves the creation of artificial datasets that retain some the statistical properties of real genomic information without compromising individual privacy [21]. SD, generated through sophisticated generative models, mirrors some of the the underlying patterns and relationships within genomic datasets [30]. This allows researchers to conduct analyses, develop predictive models and validate hypotheses in a manner that complies with the relevant data protection regulations [7].

Beyond privacy protection, the need for SD in genomic research arises from several fundamental challenges associated with real-world datasets. Genomic and clinical data are often scarce, expensive to acquire, and subject to strict data-sharing regulations, limiting their availability for large-scale studies. Additionally, biases in genomic datasets—such as underrepresentation of minority populations or imbalance in cancer subtypes—can impact model performance and generalization. SD helps mitigate these issues by offering privacy-preserving, statistically representative datasets that augment real data, improve model robustness, and facilitate algorithm validation in controlled settings.

The scientific community expects SD to play an increasingly pivotal role in genomic research by enabling more reproducible studies, supporting machine learning applications, and addressing data limitations in AI-driven diagnostics. As generative models advance, synthetic datasets have the potential to enhance precision medicine by providing diverse, high-quality training data that better reflect real-world biological variability.

One of the most common applications of SD in genomics is the generation of synthetic cancer genomic datasets, which are widely used to validate algorithms for cancer genomics analysis. These synthetic datasets enable researchers to test and refine computational methods for identifying genetic variants, tumor subtypes, and potential biomarkers while mitigating challenges related to data availability, privacy concerns, and class imbalance in real-world datasets. This approach guarantees the continued security of sensitive patient data while facilitating the unimpeded continuation of high-quality research.

The generation of SD not only safeguards the privacy of research participants but also has the potential to significantly accelerate collaborative genomic research as discussed in the stufy of Li et al. [19]. The safe and ethical sharing of representative data facilitated by synthetic datasets enables multi-institutional studies and fosters a culture of innovation in computational oncology [6]. This approach enhances the reproducibility and generalisability of findings across diverse populations, thereby facilitating more robust and inclusive cancer research.

Furthermore, SD enables the testing and training of machine learning algorithms, which are becoming increasingly pivotal in modern oncology for tasks ranging from predictive modelling to decision support [12, 17]. Unlike real-world data, which can contain missing values, inconsistencies, and potential biases, SD provides a controlled

environment for developing and validating machine learning models. A key advantage is that synthetic datasets can be tailored to specific research needs, balancing rare and frequent cases to improve classification performance. This controlled generation of training data improves model generalization, making AI-based diagnostics more robust in real-world applications.

The use of artificial intelligence techniques, particularly deep learning and machine learning, has transformed the way SD is generated in oncological genomics, enabling the creation of datasets that preserve some of the statistical characteristics of real data, supporting research in conditions of limited clinical data while ensuring privacy protection.

However, although these developments represent a significant shift in the field, several challenges still need to be addressed, including noise control in generated data, the accuracy of simulations, and their validation against real data. More broadly, using synthetic data for a particular data analysis or machine learning task requires understanding what statistical characteristics of the data are required to be preserved for that task, and whether the given SD algorithm preserves these statistical characteristics of the data. This is especially important in a field like cancer genomics where many important observed phenotypes are the result of the interactions of many smaller relatively weak interactions, which are not typically generated by SD algorithms.

The increasing number of scientific publications and the adoption of these tools in progressively larger studies indicate that SD generation is having a substantial impact on oncological research. Nevertheless, consolidating these technologies and applying them on a large scale require further studies to enhance the quality and reliability of SD, making them increasingly integrable into research processes [8].

The main objective of this work is to understand how global scientific interest in SD generation has evolved over the past ten years, from 2013 to 2024. The analysis is based on an extensive bibliometric study of data extracted from SCOPUS [3], one of the primary academic platforms used by researchers to identify and study relevant articles and studies. Through this analysis, the aim is to provide a clear picture of how the scientific community has embraced artificial intelligence as a resource to generate SD, not only to compensate for the scarcity of real data but also to explore new horizons in oncology research.

A central aspect of the article is the identification of the most significant applications of these techniques. The main research focus is multi-omics analysis, which aims to study the complex interactions among genomic, transcriptomic, and proteomic data to understand the biological mechanisms underlying cancer. In this context, multi-omics data integration serves as a key methodological approach, facilitating the combination of heterogeneous datasets to improve the robustness and interpretability of multi-omics analysis.

Another key area is the analysis of genomic heterogeneity in tumors, which enables the classification of tumors into molecular subtypes, facilitating the development of personalized therapies.

The work also explores the main algorithms used for generating this data, highlighting how machine learning and deep learning are playing a crucial role. However, challenges also emerge, such as managing noise in the data and optimizing algorithmic performance.

The structure of the paper is as follows: The Sect. 2 outlines the criteria for selecting the studies analyzed, detailing the bibliometric approach used and the data sources, primarily extracted from SCOPUS. It also clarifies the methods of analysis and the inclusion/exclusion criteria for scientific works. The Sect. 3 systematically addresses the eight key research questions underpinning the study: (1) the number of studies published on genomic data generation in oncology, (2) the main publication channels, (3) the countries with the most active research, (4) the most commonly used applications and methods, (5) the algorithms frequently employed, (6) the datasets most used, (7) the parameters for performance evaluation, and (8) the main challenges encountered. Each question is explored with detailed analyses, graphs, and tables to visually represent the findings. The Sect. 4 explores emerging trends, identifying prevalent research themes and their evolution over the past ten years. The final Sect. 5 summarizes the main findings, reflects on ongoing challenges, and offers insights into future directions for SD a generation in genomic oncology research.

2 Research methodology

This section defines the research methodology used in this study, structured into research questions. The research questions are as follows:

- *RQ1:* How many scientific studies on the generation of genomic data in cancer research using artificial intelligence have been published between 2013 and 2024?
- *RQ2:* What were the most significant publication channels?
- *RQ3:* Which countries had the most active research centers in this field?
- *RQ4:* What were the main applications and methodologies used in these studies?
- *RQ5:* What were the most commonly used algorithms for generating synthetic data?
- *RQ6:* Which were the most commonly used datasets?
- *RQ7:* What parameters were used to assess performance?
- *RQ8:* What are the main research challenges specifically related to the generation and use of synthetic data in genomics?

To achieve our goals, Elsevier's multidisciplinary bibliographic database SCOPUS served as the primary source for data extraction. This comprehensive academic database serves as a central platform for researchers seeking a nuanced understanding of the academic landscape, providing details and academic insights and supporting them in conducting citation analysis, identifying trends, and assessing the relevance of scholarly articles [24].

To limit the scope of our research, we used the following search string: ALL (("artificial intelligence" OR "machine learning" OR "deep learning" OR "reinforcement learning") AND ("synthetic data" OR "generated data" OR "simulated data") AND genomic* AND (cancer* OR tumor* OR neoplasia*) AND NOT image AND NOT imaging AND NOT radiology AND NOT mri AND NOT ct AND NOT x-ray) AND PUBYEAR> 2013 AND PUBYEAR< 2025 AND (LIMIT-TO (DOCTYPE, "ar") OR LIMIT-TO (DOCTYPE, "cp")) AND (LIMIT-TO (EXACTKEYWORD, "Human") OR LIMIT-TO (EXACTKEYWORD, "Humans")) AND (LIMIT-TO (LANGUAGE, "English")).

To refine the results of our analysis, we used the following inclusion and exclusion criteria:

- *Inclusion criteria:* The articles eligible for analysis were limited to those: written in English; written between 2013 and 2024; filtered by human and humans keywords; they must be either a journal article or a conference paper.
- *Exclusion criteria*: In the case of duplicate articles, less recent versions were not included in the analysis.

Our analysis returned 257 results in Scopus, which 198 (77%) were original investigative studies, while 59 (23%) were review articles. This distinction highlights that the majority of research in SD generation for genomics focuses on novel methodological developments and experimental validation, while reviews primarily discuss broader applications, challenges, and future perspectives.

Furthermore, of the 257 studies analyzed, 64 (25%) explicitly reported the use of clinical data alongside genomic data in AI/ML-based analyses. These studies integrated patient demographics, clinical outcomes, and treatment responses with synthetic genomic datasets to improve model accuracy and enhance translational research applications in oncology.

We then examined the more common subject areas, as shown in Fig. 1 the main areas involved are *"Biochemistry, Genetics and Molecular Biology"* with 177 papers, *"Computer Science"* with 106 papers, *Mathematics* with 100 and *"Medicine"* with 82 papers.

3 Research questions

All 257 articles have been used to answer RQ1, RQ2 and RQ3. The list of articles was based on the search strings outlined in the Research Methodology.

3.1 RQ1: How many scientific studies on the generation of genomic data in cancer research using artificial intelligence have been published between 2013 and 2024?

This research question aims to quantify the interest of the international scientific community in the generation of genomic data using artificial intelligence for the study of cancer over the last 10 years. As shown in Fig. 2, the number of publications remained relatively low until 2018, with the number of publications each year less than or equal to 20 (15 in 2014, 14 in 2015, 20 in 2016, 19 in 2017, and 20 in 2018).

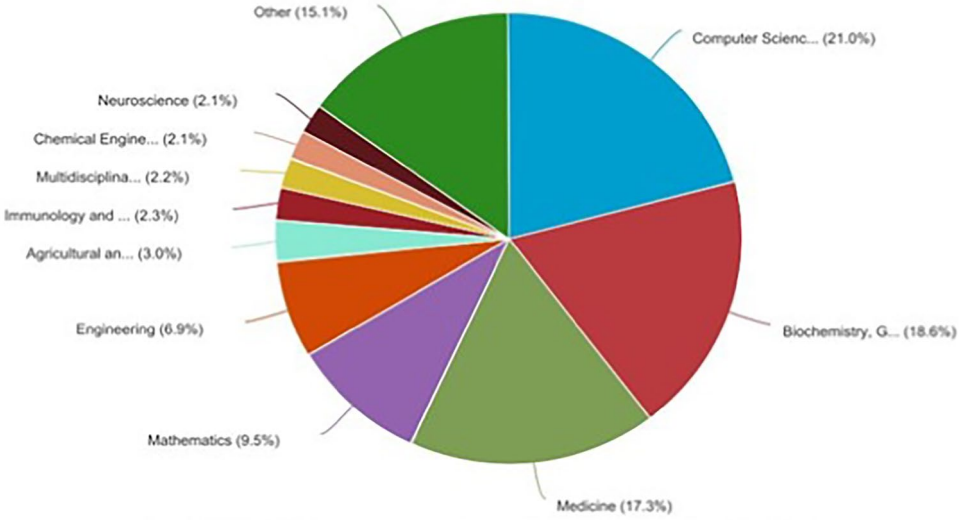

Fig. 1 Documents by subject area for the Scopus database

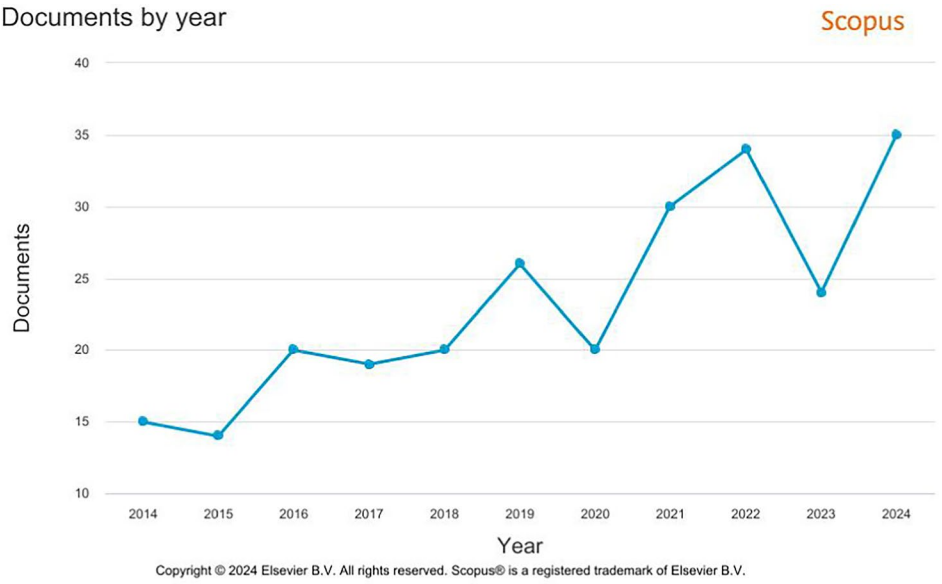

Fig. 2 Number of documents per year from 2014 to 2024

Since 2019, there has been a rapid growth in interest in this topic, with an annual growth rate of 8.84%, reaching 34 publications in 2022 and 35 in 2024, demonstrating a growing interest in this topic in recent years.

To analyze this trend in detail, the period has been divided into two blocks: 2013–2018 and 2019–2024. During the first block, the average number of publications remained relatively stable, reflecting the limited application of synthetic data in cancer genomics. However, in the second block (2019–2024), the field witnessed a notable increase in publications, probably due to advances in AI technologies and the growing recognition of SD as a valuable resource in genomic oncology research. From 2014 to October 2024, 246 articles and 11 conference papers were published; the average age of the documents was 4.21 years, and the average number of citations per document was 20.63.

The increased publication rate since 2019 highlights the role of artificial intelligence in addressing the scarcity of real clinical data by generating synthetic datasets that simulate complex biological scenarios. This trend suggests that as AI models, especially deep learning, continue to evolve, the generation of SD will become even more central to advancing genomic cancer research.

3.2 RQ2: What were the main publication channels?

With this research question, our goal is to identify and explore the primary platforms through which studies on SD generation in genomic cancer research are published. As shown in Fig. 3, academic journals have been the most prominent channels, reflecting the importance of peer-reviewed dissemination in this field. Leading journals include "BMC Bioinformatics," "Bioinformatics," and "PLOS ONE," which are well-regarded for their focus on bioinformatics and computational biology. These journals play a crucial role in this domain, likely due to their visibility and relevance in publishing interdisciplinary studies that bridge artificial intelligence, genomics, and oncology. In addition, the open access options provided by these journals improve the global reach of research findings, enabling wider access to the latest developments in SD applications.

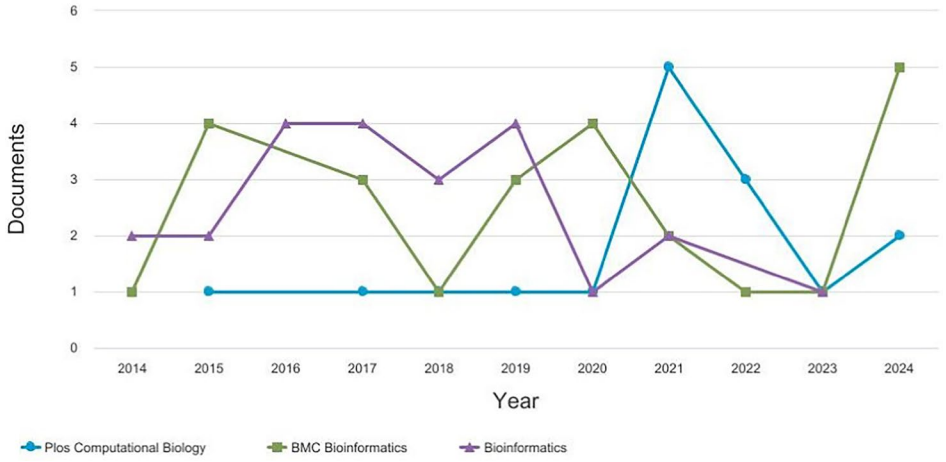

Fig. 3 Distribution of documents by year by source, from 2014 to 2024

Table 1 Key journals publishing research on synthetic genomic data in cancer, each with at least three articles

Journal	Number of publications
BMC Bioinformatics	25
Bioinformatics	23
PLOS Computational Biology	16
PLOS ONE	14
Statistics in Medicine	7
Briefings in Bioinformatics	6
IEEE/ACM Transactions on Computational Biology and Bioinformatics	6
Scientific Reports	6
BMC Genomics	5
Computers in Biology and Medicine	5
Genome Biology	5
Nucleic Acids Research	5
Frontiers in Genetics	4
Genome Research	4
Journal of Biomedical Informatics	4
Journal of Computational Biology	4
Nature Communications	4
Artificial Intelligence in Medicine	3
BMC Medical Genomics	3
BMC Medical Informatics and Decision Making	3
BMC Research Notes	3
Biometrics	3
Cell Systems	3
Statistical Applications in Genetics and Molecular Biology	3

Table 1 presents a summary of the main journals that our research identified as key platforms for studies on SD in cancer genomics, each with at least three publications on the topic. This list underscores the diversity of outlets that actively support the dissemination of work in this growing field, highlighting journals across both the bioinformatics and medical informatics disciplines.

In recent years, there has been a clear increase in publications within these leading journals, particularly since 2019. This trend points to the growing interest in SD as AI technologies advance, with an increasing number of journals embracing this

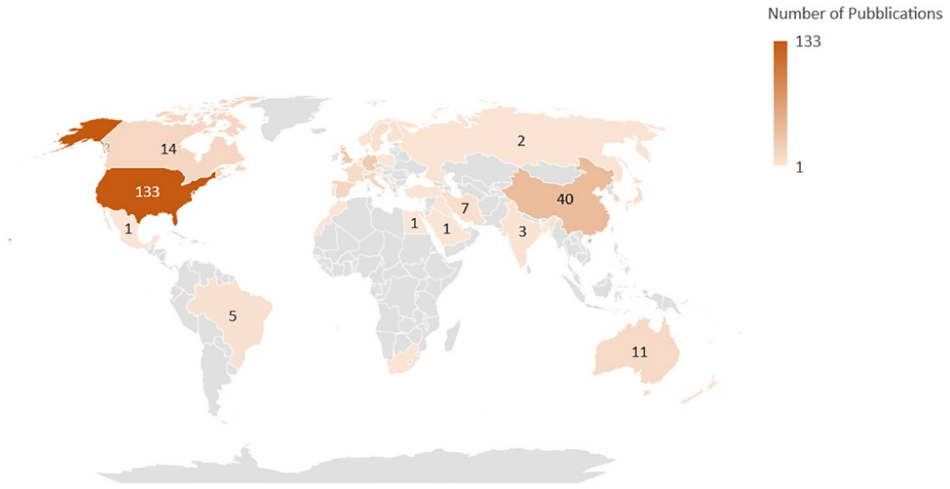

Fig. 4 Distribution of publications by country

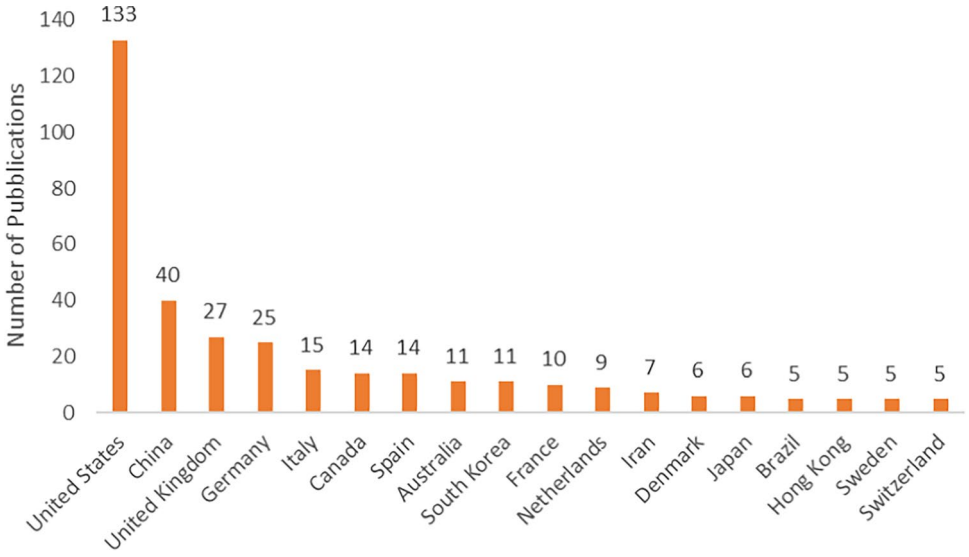

Fig. 5 Number of articles published by country from 2014 to 2024

interdisciplinary field. Advances in machine learning techniques like deep learning have made it possible to generate complex synthetic genomic data, which expands the role of these journals as platforms for innovative research.

Moreover, a significant portion of these publications is open access, facilitating greater reach and accessibility, especially for researchers in regions with limited access to subscription-based journals. Open access is crucial in this rapidly evolving field, as it ensures that the latest findings are available to the widest possible audience, fostering collaboration and accelerating progress.

3.2.1 RQ3: Which countries are most active in this research?

This research question examines which countries have made the most significant contributions to the study of SD in genomics and oncology. Figures 4 and 5 show that the largest number of publications come from the United States (133 articles), followed by China (40 articles) and the United Kingdom (27 articles). Figure 6 illustrates publication

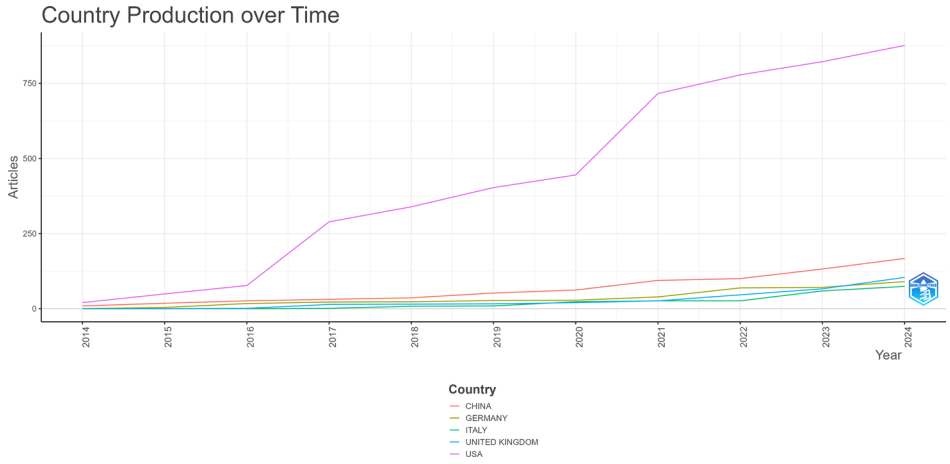

Fig. 6 Publication trends from the top five countries, 2014 to 2024

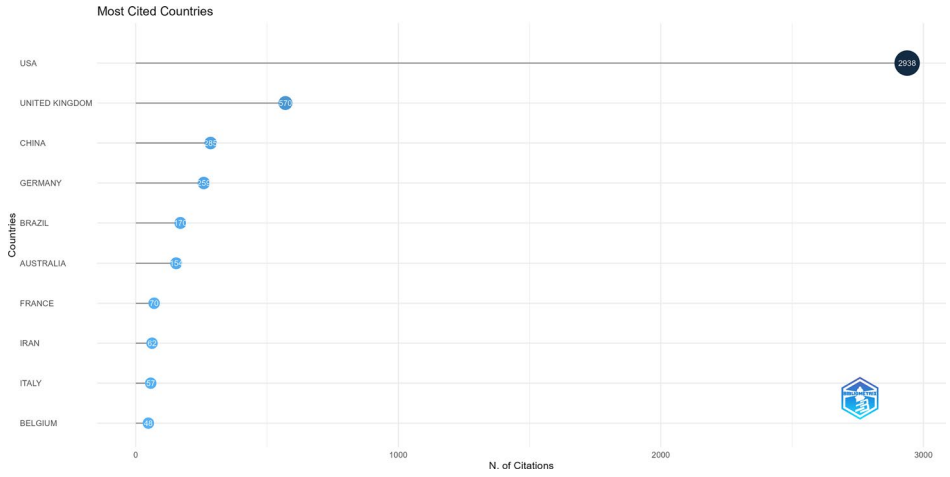

Fig. 7 Top ten most-cited countries from 2014 to 2024

trends from 2014 to 2024 for the top five countries, while Fig. 7 highlights the ten countries with the highest citation counts over the same period.

Figure 8 shows the countries of origin for the primary authors in the top 20 countries and their levels of international collaboration. Countries with a multiple countries publication (MCP) rate above 50%—such as Germany, Italy, Iran, Spain, and Belgium—are leaders in international collaboration, while Mexico, Switzerland, and the Netherlands have a higher rate of single-country publications (SCP).

The collaboration network map (Fig. 9) visualizes connections between countries in this research field, where the circle size represents the number of publications and line thickness indicates collaboration strength. The United States, China, and Germany show the strongest connections, forming four main clusters based on the frequency of co-authored publications.

3.2.2 RQ4: Which applications and methods have been used most?

To investigate the application domains and AI techniques used in SD Generation in Genomics Cancer Medicine, we conducted a clustering analysis of the titles and abstracts of scientific articles. This approach allowed us to identify and distinguish

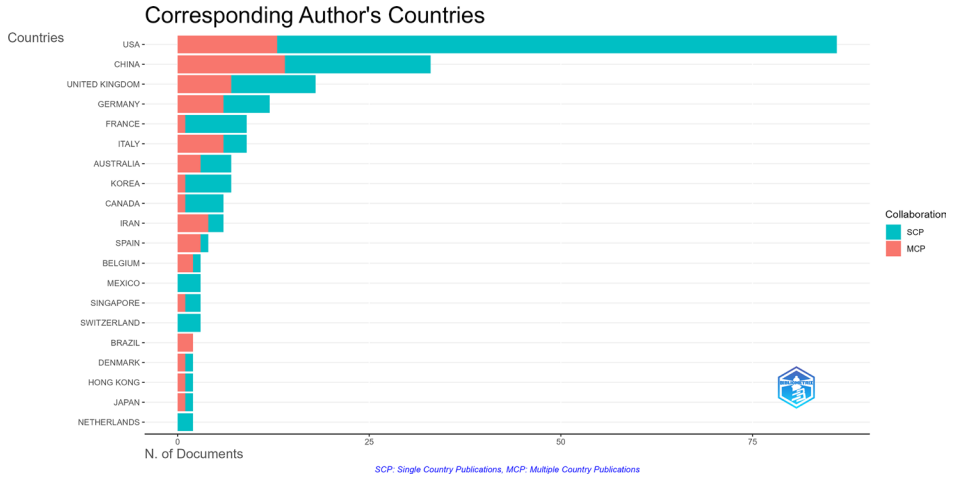

Fig. 8 Corresponding authors' countries and collaboration levels

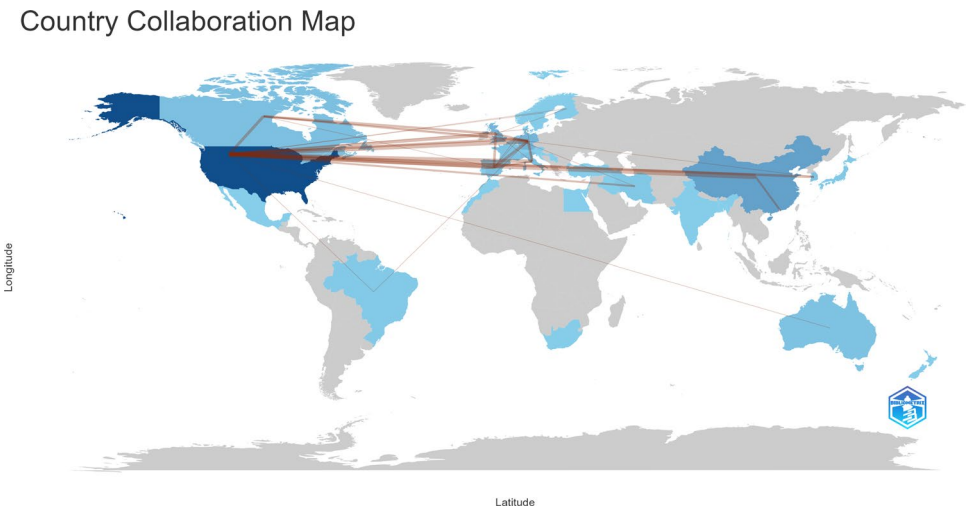

Fig. 9 VOSviewer map of international co-authorship. Of 49 countries, 18 have at least five publications, forming four main clusters

potential thematic clusters within our dataset, providing a deeper understanding of the emerging trends and applications in this field. The process was divided into following stages:

- *Data preprocessing:* Before performing the clustering analysis, the titles and abstracts of the articles were preprocessed. This step included the removal of stopwords, conversion to lowercase letters, and text normalization to reduce lexical complexity.
- *Text clustering:* After preprocessing, a K-Means clustering algorithm was applied. A model based on text similarity was used to group the articles by common themes, represented by the clusters in the graph (Fig. 10).

The clustering analysis of the articles revealed the following distribution: *Cluster 3* contains the highest number of articles, with 90 contributions. *Cluster 2* follows with 57 articles, while *Cluster 1* includes 48 articles. These three clusters represent the majority of the articles analyzed, suggesting the presence of central topics within the dataset, likely related to emerging or well-established research fields.

Fig. 10 The graph showing the distribution of articles in the 5 clusters

Cluster 0, with 37 articles, is slightly less populated but still represents a significant portion of the dataset. Finally, *Cluster 4* is the smallest, with only 25 articles, indicating that these studies address more specific or less common topics compared to the other groups.

To improve the clarity of our categorization, we refined the definition of each cluster based on the specific research problems addressed and the role of SD in supporting these studies. While all clusters involve SD generation, their focus varies depending on the applications and methodologies used:

- *Cluster 3: Multi-Omics Data Integration and Machine Learning*. Cluster 3 is the largest with 90 articles, primarily gathering studies focused on the integration of multi-omics data and the application of machine learning techniques to analyze biological complexity. These studies explore approaches for integrating data from genomics, transcriptomics, and proteomics, aiming to identify patterns and relationships in large-scale datasets. This cluster represents one of the most active research areas, where advanced machine learning models play a crucial role in analyzing large biological datasets, helping to better understand the complex mechanisms behind diseases.
- *Cluster 2: Deep Learning in Genomic Applications*. Cluster 2, with 57 articles, focuses on deep learning methodologies applied to genomics and epidemiology. The studies in this group often explore the use of deep neural networks for analyzing genomic data, including identifying genetic variants, predicting gene expression patterns, and uncovering functional genomic elements. These methods enable advanced data interpretation rather than directly performing genome sequencing. This field

is highly innovative, as deep learning technologies allow researchers to uncover significant insights from vast genomic datasets, speeding up the discovery of new biomarkers and therapeutic targets.

- *Cluster 1: Tumor Subtyping and Genomic Heterogeneity*. Cluster 1 contains 48 articles. These studies examine the different mutations and genetic variations present in tumors, with a focus on molecular subtyping techniques. The main goal is to use SD to enhance the classification of tumor subtypes, aiding in the identification of genetic biomarkers and personalized therapeutic strategies. This area of research is critical for developing personalized therapies and improving cancer treatment, as understanding these genetic variations helps identify specific patient subgroups that might respond better to certain drugs.
- *Cluster 0: Genetic Optimization and Gene Expression Network Analysis*. Cluster 0 with 37 articles, is the smallest group but includes studies that explore highly specialized topics. These articles delve into genetic parameter optimization, transcriptomic data correlations, and computational analysis of gene expression networks. SD plays a role in improving the accuracy and scalability of models used to infer gene regulatory networks and optimize genetic feature selection. This suggests that this cluster represents a more niche area of research, with applications that may not be as widely spread but are still important for advancing biological modeling and computational techniques.
- *Cluster 4: Synthetic Data Generation for Medical Research*. Cluster 4, with 25 articles, includes studies focused on generating SD, particularly in the field of hematology. Unlike other clusters, which incorporate synthetic data as part of broader methodologies, this group specifically investigates the creation, validation, and application of synthetic datasets for medical research. These articles explore advanced methods for creating artificial data using artificial intelligence, with the aim of accelerating medical research and precision medicine. SD generation is especially valuable for simulating complex scenarios or filling in gaps in clinical data, allowing researchers to test new hypotheses without the need for real patient data.

From this point forward, we will focus on the two most cited articles within each cluster. This choice allows us to examine the contributions that have had the greatest impact on the scientific community, as the number of citations is often a good indicator of relevance and influence. By analyzing the most cited works, we can highlight the methodologies and approaches that have been recognized as the most significant in their field. This way, we reduce the complexity of the dataset while still maintaining a focus on what matters most for our study.

3.2.3 RQ5: Which were the most commonly used algorithms?

This section analyzed the frequency and use of algorithms in the context of the research conducted in this article. Algorithms are fundamental tools for advancing gene regulatory network inference, making their identification crucial for understanding methodological trends and areas of innovation. This focused review aimed to highlight the most commonly used algorithms, emphasize their applications, shown in Table 2, and discuss their importance within the current landscape of genomic data analysis. To provide a clearer understanding of their relevance, we have included a quantitative assessment, indicating the number of papers within each cluster that employ each algorithm. This

Table 2 Overview of commonly used algorithms in genomic data analysis across different clusters

Cluster	Study	Algorithm(s)	Description and key points	Number of papers using this algorithm
Cluster 0	[5]	PIDC (partial information decomposition and context)	Used to infer gene regulatory networks from single-cell transcriptomic data, outperformed methods like MI and ARACNE by capturing higher-order dependencies	5
	[23]	MEGENA (multiscale embedded gene co-expression network analysis)	Designed for large-scale gene co-expression networks, effective in detecting biologically relevant clusters across scales, surpassing PMFG and WGCNA in managing large networks and computational efficiency	3
Cluster 1	[20]	CancerDetector and CancerLocator	CancerDetector improves accuracy in distinguishing tumor cfDNA using α-value for methylation analysis, surpassing CancerLocator in sensitivity and precision, especially at low tumor fractions	4
	[9]	SPhyR	Tumor phylogeny estimation using k-Dollo parsimony model, improving precision by addressing sequencing errors; more flexible than SCITE for complex tumor evolution	6
Cluster 2	[32]	CNNs, RNNs, Autoencoders	Deep learning models for pattern recognition in genomic data, including CNNs for sequence patterns and RNNs for sequential analysis. Autoencoders used for data dimensionality reduction	8
	[11]	LIMITS (learning interactions from microbial time series)	Sparse linear regression with bootstrap aggregation to infer microbial interactions, managing noise in metagenomic data effectively	3
Cluster 3	[29]	NMF (non-negative matrix factorization), jNMF, iNMF	Used to integrate heterogeneous genomic data, with iNMF addressing limitations of jNMF in handling source-specific noise	7
	[27]	Precision Lasso	Advanced Lasso variant that considers variable correlations and dependencies, enabling stable selection in genomic data with complex relationships	5
Cluster 4	[14]	Neural networks (CNNs, RNNs), clustering (k-means, DBSCAN), regression, ensemble models	Machine learning techniques applied to COVID-19 data for disease progression prediction and patient clustering based on clinical profiles	6
	[22]	Deep neural networks, latent variable models, Incremental learning	Deep learning and clustering for patient stratification, with incremental learning to adapt models to real-time health data from wearables	4

addition offers a more structured perspective on the prevalence and role of these methodologies across different research domains.

As for Cluster 0 we have: Chan et al. [5] and Song and Zhang [23] and specifically:

- The study conducted by Chan et al. [5] focused on developing and validating the PIDC algorithm (Partial Information Decomposition and Context), designed to infer gene regulatory networks from single-cell transcriptomic data. This algorithm was compared to other common methods, such as Mutual Information (MI) and ARACNE, and showed superior performance due to its ability to capture higher-order dependencies between genes. The PIDC approach included a detailed analysis of gene-to-gene interactions and outperformed existing algorithms like CLR (Context Likelihood of Relatedness) in several key network inference tasks.

- The study conducted by Song and Zhang [23] focused on developing and validating the MEGENA (Multiscale Embedded Gene Co-expression Network Analysis) algorithm, designed to build large-scale gene co-expression networks and identify clustering structures across multiple levels. This algorithm was compared to existing methods like the Planar Maximally Filtered Graph (PMFG) and Weighted Gene Co-expression Network Analysis (WGCNA), showing superior performance in handling large networks and reducing computational complexity. MEGENA stood out for its ability to detect biologically relevant clusters at different scales, outperforming algorithms like ARACNE and CLR (Context Likelihood of Relatedness) in key tasks for gene network inference, thanks to its precision in identifying meaningful gene-gene interactions.

For Cluster 1 we have Li et al. [20] and El-Kebir [9] and specifically:

- In the study conducted by Li et al. [20], two main algorithms are discussed for the detection of circulating tumor DNA (cfDNA) through DNA methylation analysis. One of them is CancerDetector, which focuses on detecting tumor cfDNA by analyzing the methylation profiles of adjacent CpG sites in a single cfDNA strand. The goal is to probabilistically model the joint methylation states to exploit the pervasive nature of methylation. The algorithm introduces the concept of the α-value, which represents the average methylation states of all CpG sites in a sequencing read, allowing for a more accurate distinction between tumor and normal cfDNA compared to traditional methods using the β-value (which averages the methylation of individual CpG sites across all reads). This approach enables the detection of small amounts of tumor cfDNA, even at low sequencing coverage (from 1x to 10x), with increasing accuracy as coverage increases. The other algorithm is CancerLocator, which is based on the traditional β-value to estimate the tumor cfDNA fraction. The method combines the average methylation values of specific tumor markers to estimate the tumor cfDNA fraction. However, compared to CancerDetector, CancerLocator shows inferior performance, particularly in terms of sensitivity and precision when detecting low tumor fractions.
- In this study conducted [9], an innovative approach for tumor phylogeny estimation from single-cell sequencing data is presented, focusing on the SPhyR algorithm. This method employs the k-Dollo parsimony model, which allows mutations to be gained only once but lost multiple times, an essential feature for accurately describing tumor evolution. Tumors, in fact, are characterized by significant genetic heterogeneity, where mutations can be lost due to copy number aberrations, a phenomenon not considered in simpler evolutionary models. SPhyR also addresses common issues related to single-cell sequencing errors, such as false positives and false negatives, correcting these errors to achieve a more precise phylogenetic representation. A comparison is made with the SCITE algorithm, which is based on the "infinite sites" model. This model is more rigid compared to the SPhyR model, as it does not allow for either the loss or recurrence of mutations, limiting its applicability to tumors where the loss of genomic regions is frequent. SCITE, in fact, assumes that each mutation occurs only once and remains stable over evolutionary time, a premise that does not fit well with the complexity of tumor evolution.

For Cluster 2 we have Zou et al. [32] and Fisher and Mehta [11] and specifically:

- In the study conducted by Zou et al. [32], it explores how deep learning algorithms can transform the field of genomics, addressing complex challenges such as predicting the effects of genetic variants and understanding the mechanisms that regulate gene expression. The study focuses on how convolutional neural networks (CNNs), recurrent neural networks (RNNs), and feed-forward networks can be valuable tools for identifying hidden patterns in large volumes of genetic data. The use of CNNs has enabled the detection of repeated patterns within DNA sequences, allowing for the identification of regulatory regions, such as enhancers, which play a crucial role in gene activation. RNNs, on the other hand, have been fundamental in analyzing sequential data, being able to "remember" past information within the DNA sequence to predict future events, such as splicing mechanisms or temporal changes in specific genetic variants. Autoencoders were also used, making it possible to reduce the size of massive genomic datasets, allowing for a clearer and more accessible picture of how variations in the data can be correlated with specific genetic traits or diseases.
- In the study conducted by Fisher and Mehta [11], particular attention was given to analyzing the ecological interactions between microbial species in the human microbiome. The researchers used advanced methods like sparse linear regression and bootstrap aggregation. These techniques were developed to overcome common challenges in inferring species interactions, such as the difficulty in drawing reliable conclusions from metagenomic data, since measured relative abundances don't always reflect real correlations between species. The study introduces a new approach called LIMITS (Learning Interactions from Microbial Time Series), which allows for the inference of a discrete Lotka-Volterra model for microbial dynamics, based on time series data. This algorithm combines sparse linear regression with bootstrap aggregation to stabilize estimates and reduce the impact of experimental noise, which is common in this type of analysis. What makes LIMITS stand out is its ability to infer the topology of ecological interactions not only from absolute abundances but also from relative abundances, thereby overcoming the sum constraint on abundances that often complicates the inversion of regression matrices. The algorithm uses a stepwise regression selection process, where species interactions are added to the model one by one, based on improvements in the model's predictive power. To enhance the stability of the solution, bootstrap aggregation is then applied: this technique, by randomly partitioning the data and taking the median of the estimates, preserves the sparsity of the interaction matrix, meaning it only includes the most significant species interactions.

For Cluster 3 we have Yang and Michailidis [29] and Wang et al. [27] and specifically:

- The study conducted by Yang and Michailidis [29] introduced a new methodology based on Non-Negative Matrix Factorization (NMF) for analyzing heterogeneous multi-modal omics data. In their research, the authors focused on developing methods to integrate various genomic data sources, addressing one of the main challenges in this field: the heterogeneity of the data. NMF, already widely used for high-dimensional data analysis, offers an intuitive biological interpretation by imposing a non-negativity constraint, allowing for the decomposition of data into modules that represent key biological signals. The authors further expanded

this methodology by introducing joint NMF (jNMF), which integrates multiple datasets, such as gene expression, DNA methylation, and microRNAs, and detects coordinated activity across these sources. However, while jNMF is effective at identifying homogeneous effects, it showed limitations when dealing with source-specific noise, making it less robust in the presence of heterogeneous data. To address these limitations, Yang and Michailidis developed integrative NMF (iNMF), an algorithm that distinguishes between homogeneous and heterogeneous effects, significantly enhancing the ability to identify relevant biological modules even in highly heterogeneous conditions.

- The study conducted by Wang et al. [27] introduces the Precision Lasso, a new algorithm designed to overcome the limitations of traditional Lasso in variable selection, particularly when working with complex genomic data that features strong correlations and linear dependencies between variables. Traditional Lasso, while widely used for reducing the dimensionality of regression models, tends to randomly select one variable among those that are highly correlated, often overlooking other variables that may hold equal statistical importance. Moreover, when there are linear dependencies between explanatory variables, the standard Lasso might combine them in ways that are not necessarily biologically meaningful. To address these issues, Precision Lasso introduces a regularization technique that accounts not only for the correlation between variables but also for their linear dependencies by using covariance and inverse covariance matrices. This approach enables more stable and consistent identification of the most relevant variables.

Finally, for Cluster 4 we have Haendel et al. [14] and Shameer et al. [22] and specifically:

- In the study conducted by Haendel et al. [14], various types of machine learning algorithms, such as neural networks and clustering techniques, were used to process and analyze clinical data related to the COVID-19 pandemic. Neural networks, in particular, were employed for their ability to model complex relationships between variables, such as patient clinical characteristics and disease outcomes. Convolutional neural networks (CNNs), often used for image processing, were adapted for analyzing tabular and temporal data, leveraging their ability to identify patterns in real-time clinical data. Recurrent neural networks (RNNs), especially variants like Long Short-Term Memory (LSTM) networks, were used to model temporal sequences in clinical data, such as tracking patient condition during hospitalization, the progression of infections, and the effectiveness of treatments. This approach enabled the prediction of disease progression for individual patients, helping to identify those at risk of severe complications like acute respiratory distress syndrome or other systemic conditions. Another widely used algorithm in the study was clustering, specifically methods like k-means and DBSCAN (Density-Based Spatial Clustering of Applications with Noise). These algorithms were employed to group patients into homogeneous cohorts, identifying similarities in treatment responses and risk profiles. The k-means algorithm was used to segment patients based on variables such as age, comorbidities, and laboratory test results, while DBSCAN helped identify subgroups of patients with atypical characteristics or rare complications. For predictive analysis, linear and logistic regression algorithms were used to model relationships between clinical variables (such as oxygen levels,

body mass index, and lab test results) and clinical outcomes, such as the need for mechanical ventilation or the risk of mortality. The authors also applied more advanced techniques like random forest and gradient boosting to improve model accuracy and manage data complexity, particularly when dealing with non-linear interactions between variables. Finally, the use of ensemble learning models, such as deep neural networks combined with boosting methods, was explored to enhance the robustness of predictions. These approaches allowed the combination of multiple models' strengths, reducing the risk of overfitting and improving generalization on heterogeneous clinical data from different organizations.

- In the study conducted by Shameer et al. [22], several advanced machine learning and data analysis techniques are used to manage and interpret real-time biomedical and healthcare data streams. One of the key techniques is the use of deep neural networks, which are applied to model complex relationships between clinical, genomic, and other biomedical data. These deep neural networks have proven highly effective in handling large-scale, high-dimensional datasets, such as those generated from omics profiling, including genomics, proteomics, and metabolomics. Variants of neural networks, like convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are specifically used to recognize patterns in heterogeneous data, playing a crucial role in areas such as medical imaging diagnostics and the prediction of clinical events. Another technique employed in the study is latent variable models, such as Latent Dirichlet Allocation and mixture models. These are used to uncover hidden variables that explain data variability. These models are especially useful for simplifying complex datasets, creating compact representations that retain the essential characteristics of the original data. Clustering is also a valuable technique, grouping data based on shared features to discover new subgroups of patients with similar, but previously unrecognized, phenotypes. This approach has been particularly helpful in identifying subpopulations of patients with specific genomic or clinical profiles, such as in the stratification of type 2 diabetes patients, allowing for personalized therapies tailored to the characteristics of each group. Finally, one of the key aspects addressed in the study is the use of incremental learning algorithms, which enable predictive models to be continuously updated with new real-time data. This is especially relevant when integrating data from wearable devices or health sensors, where the constant stream of information must be processed and incorporated into models to maintain their accuracy and effectiveness over time.

In addition to the methodologies already discussed in Table 4, we have identified other relevant techniques for SD generation and genomic analysis. While some studies are among the most cited in their respective clusters, it is crucial to also include methodologies that, despite being less frequently referenced, play a significant role in computational oncology research.

Below, we provide a list of methodologies not previously mentioned but equally essential in this field:

For inference networks:

- *Non-negative matrix factorization (NMF)*—Used for multi-omics data integration and dimensionality reduction, improving model interpretability.

- *Precision Lasso*—Applied for feature selection in high-dimensional genomic datasets, reducing computational complexity while preserving critical information.

For tumor evolution models:

- *SPhyR*—Used for tumor phylogeny reconstruction, facilitating the analysis of cancer evolution and mutation patterns.
- *k-Dollo Parsimony Model*—Applied to infer tumor progression and identify relationships among different evolutionary stages of the disease.

For genetic network inference methods:

- *PIDC (partial information decomposition and context)* —A technique for inferring gene regulatory networks based on single-cell transcriptomic data, enhancing the understanding of gene interactions.
- *MEGENA (multiscale embedded gene co-expression network analysis)*—Used for identifying multi-scale gene co-expression networks, allowing for a more accurate modeling of biological complexity.

These methodologies, although not always the most cited, are fundamental for addressing specific challenges in cancer research. For example, *NMF and Precision Lasso* are crucial for feature selection in high-dimensional datasets; *SPhyR and the k-Dollo model* allow the study of tumor progression in an evolutionary context; while *PIDC and MEGENA* provide advanced tools for gene network inference and coexpression analysis, improving the biological interpretability of SD.

3.2.4 RQ6: Which were the most commonly used datasets?

This section aimed to highlight the most commonly used datasets and discuss their importance within the current landscape of genomic data analysis.

In Chan et al. [5], the authors analyzed data from three distinct experimental datasets. These include transcriptomic data from megakaryocyte-erythroid progenitors in the context of human blood cell development, early stages of mouse embryonic development, and the formation of hematopoietic cells. Through these examples, the article showcases the practical application of PIDC to uncover relationships between genes across different biological processes.

In Song and Zhang [23], the datasets used include gene expression data from The Cancer Genome Atlas (TCGA) that specifically focus on breast carcinoma (BRCA) and lung adenocarcinoma (LUAD). These datasets were analyzed to reveal multiscale gene co-expression networks and identify novel targets related to cancer biological processes.

In Li et al. [20], the dataset used is derived from multiple sources: (a) TCGA Methylation Data: this includes methylation profiles from solid liver tumors and their matched normal tissues; (b) Whole Genome Bisulfite Sequencing (WGBS) Data: obtained from studies by Chan et al. and Sun et al. [4], this dataset comprises plasma samples from healthy individuals, HBV patients, and liver cancer patients, accessed via the European Genome-Phenome Archive (EGAS00001000566 and EGAS00001001219); (c) Newly Generated Data: the authors created high-coverage WGBS data from healthy individuals' plasma and liver cancer patients to simulate higher-coverage plasma cfDNA for more accurate cancer detection.

In El-Kebir [9] the dataset used includes simulated data generated for k-Dollo phylogeny models and real-world sequencing data from a colorectal cancer (CRC) patient. Specifically, the article references data from patient CRC1, whose information was gathered and analyzed from single-cell sequencing studies. Additionally, publicly available datasets, such as those from the works of Leung et al. [18], are mentioned as part of the validation and analysis

The article of Zou et al. [32] references various genomic datasets to illustrate the application of deep learning methods. These datasets include functional genomics data, variant calling data, and gene expression profiles, as well as regulatory genomics information. The sources often involve high-throughput sequencing techniques, such as DNase I sequencing, RNA-seq, ATAC-seq for chromatin accessibility, and ChIP-seq for studying protein-DNA interactions. Additionally, the paper utilizes publicly available datasets from large-scale genomic projects like ENCODE, which provide extensive functional and regulatory annotations of the genome.

In Fisher and Mehta [11], are utilized metagenomic time-series data from the MG-RAST database. The dataset comprises relative abundances of microbial species in the gut microbiomes of two individuals, with approximately six months of daily samples for one individual and a full year for the other. Focusing on the ten most abundant species in each subject, the authors leverage this data to infer ecological interactions and identify keystone species that play a pivotal role in shaping the structure and dynamics of the microbial community.

In Yang and Michailidis [29], are employed data from The Cancer Genome Atlas (TCGA), specifically focusing on ovarian cancer. The study integrates three distinct types of genomic data: DNA methylation, gene expression, and microRNA (miRNA) expression. These datasets are utilized to assess the proposed method's ability to identify multi-dimensional modules across patient samples, which are linked to cancer-related pathways and known ovarian cancer subtypes.

The study of Wang et al. [27], utilizes datasets from The Cancer Genome Atlas (TCGA), specifically focusing on gene expression, DNA methylation, and microRNA data from three cancer types: glioblastoma, lung cancer, and breast cancer. These large-scale datasets, containing thousands of patient samples and tens of thousands of genomic features, were employed to evaluate the Precision Lasso model's ability to address the challenges posed by correlated and linearly dependent variables in high-dimensional genomic analysis.

In Haendel et al. [14], utilizes a comprehensive dataset derived from the aggregation and harmonization of electronic health record (EHR) data from numerous clinical institutions across the United States. This dataset encompasses detailed patient-level information pertinent to COVID-19, including demographics, diagnoses, laboratory results, medications, and social history. The data, curated through the N3C Enclave, includes records from patients both with and without COVID-19, with data collection extending back to January 2018.

The article of Shameer et al. [22], draws on a variety of datasets, including electronic medical records (EMRs), genomic and multi-omic data from initiatives such as The Cancer Genome Atlas (TCGA), and real-time data from health-monitoring devices and wearable sensors. These datasets encompass physiological, genomic, and environmental data, which are integrated to advance personalized medicine and enhance clinical

decision support systems. The combination of these diverse data streams facilitates a comprehensive approach to health monitoring and precision medicine.

This analysis underscores as certain datasets, such as TCGA and GEO, are frequently utilized as reference points for validating SD methodologies and benchmarking machine learning models. To quantify their recurrence, we analyzed the datasets cited in the reviewed papers and identified those that were used in at least two studies. Table 3 summarizes these frequently used datasets, highlighting their relevance in the field of genomic data analysis and SD validation.

3.2.5 RQ7: What parameters were used to assess performance?

In the context of the research that applied network inference algorithms, several performance evaluation metrics were employed to assess the effectiveness and accuracy of these computational models. These metrics offered a comprehensive understanding of each algorithm's ability to accurately infer gene regulatory relationships and served as crucial indicators of their reliability and applicability in biological data analysis.

The most frequently used evaluation parameters are:

- *Accuracy and AUC-ROC*—Standard metrics for assessing the predictive capability of classification models.
- *Precision and recall*—Key indicators for balancing false positives and false negatives, essential for detecting rare genomic variations.
- *False discovery rate (FDR)*—Used in genomic studies to control the false positive rate in multiple association analyses.
- *Pearson correlation coefficient (PCC)*—Applied to measure similarity between real and synthetic datasets, ensuring that synthesized data maintain biological correlations.
- *Feature selection stability*—Evaluates the robustness of algorithms in selecting significant biomarkers in both synthetic and real datasets.

In Chan et al. [5], the Area Under the Precision-Recall Curve (AUPR) was chosen as the primary metric because it is particularly effective when working with biological networks, which tend to be quite "sparse." In other words, in biological networks, there are only a limited number of true connections (such as gene interactions or functional relationships between nodes) compared to all possible connections. AUPR measures how well an algorithm can distinguish between true and false connections, placing more emphasis on false positives. In contexts where true connections are rare and non-connections are the majority, this metric provides a more accurate representation of the algorithm's performance. The Area Under the ROC Curve (AUROC) was also calculated,

Table 3 Most frequently used datasets in the analyzed papers

Dataset	Description	Number of papers
TCGA (The Cancer Genome Atlas)	Large-scale genomic dataset for various cancer types	15
GEO (Gene Expression Omnibus)	Repository for functional genomic studies	10
UK Biobank	Population-scale biomedical database	7
COSMIC (Catalogue of Somatic Mutations in Cancer)	Database of somatic mutations in cancer	5
GTEx (Genotype-Tissue Expression)	Gene expression across multiple tissues	4
ICGC (International Cancer Genome Consortium)	Global collaboration on cancer genomics	3

but although it is widely used in many fields, it is less effective in sparse situations like those seen in biological networks. AUROC evaluates an algorithm's ability to distinguish between true positives and false positives, but it treats both the ability to identify true connections and non-connections with equal importance. In networks with few real connections, this can lead to misleading results: an algorithm could score high on AUROC even if it correctly predicts only a few true connections, simply because there are many more true negatives (non-connections) that balance the overall score, giving a false impression of the algorithm's effectiveness.

In Song and Zhang [23], both the Area Under the Precision-Recall Curve (AUPR) and the Area Under the ROC Curve (AUROC) were used to evaluate the performance of network inference algorithms. However, AUPR was considered a more meaningful metric, especially in the context of biological networks, which tend to be very sparse. This is because AUPR focuses more on true positives rather than false positives, which is crucial when there are far fewer real connections compared to non-connections. In biological networks, where unconnected nodes vastly outnumber connected ones, AUPR provides a more accurate reflection of an algorithm's capabilities. On the other hand, AUROC tends to overestimate performance in such scenarios, as it gives equal weight to predicting both connections and non-connections. This can artificially inflate the results for an algorithm that mostly predicts non-connections, giving a false impression of its ability to correctly identify real links. For this reason, AUPR was deemed more suitable for assessing the overall performance of the algorithms tested. Another important metric was the False Discovery Rate (FDR), evaluated at various thresholds (0.01, 0.05, 0.1). This helped us assess the algorithms' ability to identify significant interactions while balancing false positives and negatives, even under stricter thresholds. The Coefficient of Variation (CV) was used to measure the stability of the algorithms' performance across different FDR thresholds, ensuring the results were consistent regardless of the applied threshold. A low CV indicated that the algorithm consistently produced reliable results. Lastly, the algorithms were tested on simulated networks as well as validated reference networks, using datasets like those from the DREAM challenge, to compare the inferred networks with known, validated biological interactions.

In Li et al. [20], the performance of the CancerDetector and CancerLocator algorithms was evaluated using several key metrics. Sensitivity was one of the main indicators of effectiveness, measuring how well the algorithm could correctly identify true positives, which is crucial for early cancer detection. Another important parameter was specificity, which assessed the algorithm's ability to avoid false positives by correctly identifying non-cancerous samples. To provide a comprehensive view of the algorithms' performance, the ROC curve (Receiver Operating Characteristic) was used. This curve plots the relationship between the true positive rate (sensitivity) and the false positive rate (1 - specificity) at various threshold levels, illustrating the trade-offs between sensitivity and specificity. Another essential metric was the Area Under the Curve (AUC), which represents the area under the ROC curve and gives a summary measure of the algorithm's overall accuracy in distinguishing between cancerous and non-cancerous samples. The higher the AUC, the better the algorithm's performance. Additionally, the Pearson Correlation Coefficient (PCC) was used to assess how well the predicted tumor fractions matched the actual tumor fractions in simulated samples. A high PCC indicated a strong agreement between predictions and real tumor fractions. Finally, standard deviation

was used to measure the variability in sensitivity and other predictions across different experiments, providing insights into the robustness of the tested algorithms.

In El-Kebir [9] the performance of the algorithms was evaluated through several key parameters. One of the main metrics used was the false positive rate (FPR), which indicates how often the algorithm mistakenly predicted a mutation where there wasn't one. Similarly, the false negative rate (FNR) measured how often the algorithm missed a mutation that should have been detected. These two metrics were essential for assessing how accurately the algorithms were able to reconstruct the tumor's evolutionary history. Another important measure was the ancestral pair recall, which looked at how well the trees produced by the algorithms matched the simulated ones. It focused on how character states evolved along the same or different branches, helping to determine if the algorithm could recreate the relationships between mutations over time accurately.

In Zou et al. [32] the performance of the deep learning algorithms was evaluated to understand how precise and useful they were in analyzing genomic data. One of the main parameters used was accuracy, which indicates how many times the model made correct predictions compared to the total data analyzed. However, given the complex nature of genomic data, where there are often many more neutral variants than pathogenic ones, accuracy alone was not enough. Precision was crucial to understand how many of the positive predictions made by the model were actually correct. This is particularly important when dealing with rare variants, such as pathogenic ones, which can easily be "hidden" among many neutral variants. Recall, on the other hand, helped assess the model's ability to not miss any of the positive variants. In other words, recall told us how good the model was at identifying all the truly relevant cases within the dataset, without overlooking critical information.

In Fisher and Mehta [11], the performance of the LIMITS algorithm was assessed by analyzing several key parameters to understand its accuracy in inferring species interactions. One of the main parameters used was the prediction error, which represents how well the model could predict species interactions based on the available data. The algorithm relies on a linear regression model that progressively added new interactions between species. At each step, the prediction error was calculated to see if including a new interaction significantly improved the model's predictive ability. Another key parameter was the stability of the estimates, achieved through bootstrap aggregation. This process helped reduce experimental noise and ensured that the results were robust and not influenced by random data fluctuations. To further validate the model, its ability to infer the correct topology of ecological interactions was also assessed, meaning how accurately the algorithm could identify which species were actually interacting with one another. This ability was tested using both absolute and relative species abundances, making the algorithm particularly versatile and adaptable to various types of metagenomic datasets.

In Yang and Michailidis [29], the authors assessed the performance of their proposed algorithm using several key parameters to measure its effectiveness in detecting biological signals and managing heterogeneous noise. One crucial parameter was the module detection score, which evaluates the algorithm's ability to correctly identify modules of observations and variables based on common signals across data sources. This score was calculated by comparing the average signal values inside and outside the modules, normalized against the signal itself. This approach allowed the researchers to quantify

how well the observed data aligned with the actual modules. Another critical parameter was the Frobenius residual error, used to compare the differences between the original data matrix and the one approximated by the algorithm. This metric provided insight into the model's overall ability to approximate the data, while keeping the primary focus on accurately identifying the relevant biological modules. However, the Frobenius error was considered less useful in scenarios where the goal was more about discovering the modules rather than simply approximating the data. By using these metrics, the authors demonstrated how their algorithm excelled in pinpointing key biological signals, particularly in complex settings with significant heterogeneous noise, outperforming other methods focused solely on homogeneous effects.

In Wang et al. [27], the performance of Precision Lasso was evaluated using several parameters to compare its effectiveness with traditional variable selection methods. One of the main metrics used was the area under the ROC curve (AUC), which measures the model's ability to distinguish between relevant and irrelevant variables, assessing both the sensitivity and specificity of the method. In addition to AUC, other performance indicators were considered, such as the number of false positives and false negatives. These were further complemented by precision, recall, and F1 score, which are essential metrics for evaluating the model's balance between minimizing classification errors and accurately capturing the truly relevant variables. Furthermore, in real-world data applications, an additional evaluation criterion was used by comparing the genes selected by Precision Lasso with those listed in the Catalogue of Somatic Mutations in Cancer (COSMIC). This comparison helped determine how many of the selected genes were actually linked to cancer, reinforcing the biological validity of the model in an oncology context.

In Haendel et al. [14], the authors used several parameters to assess the performance of the algorithms applied to clinical data related to COVID-19. One of the key metrics was predictive accuracy, which measures the model's ability to correctly predict clinical outcomes, such as disease progression or mortality risk. Predictive accuracy was calculated by looking at the ratio of correct predictions to the total number of predictions, for both positive and negative cases. This metric helped evaluate the model's ability to generalize predictions to data that was not part of the training set. Another important parameter was accuracy, which reflects the percentage of cases correctly classified by the model. The authors compared accuracy across different models to identify which algorithms were more effective in handling complex clinical data, often filled with heterogeneous and non-linear variables. In addition to accuracy, the Area Under the ROC Curve (AUC-ROC) was used to evaluate the model's ability to distinguish between different classes, particularly for binary classification models. This metric allowed the authors to assess how well the model could differentiate between positive and negative cases. The authors also examined sensitivity and specificity, two crucial parameters for assessing the model's ability to accurately detect both positive and negative cases. Sensitivity measured the proportion of true positives correctly identified by the model, while specificity evaluated the proportion of true negatives detected. These parameters were essential for understanding the model's effectiveness across different clinical conditions. Lastly, to ensure a comprehensive evaluation of the model's performance, the authors considered both Positive Predictive Value (PPV) and Negative Predictive Value (NPV). These metrics indicate the likelihood that a subject identified as positive or negative by

the model is actually so. These indicators were especially useful for monitoring the quality of predictions in critical clinical scenarios, such as managing patients at high risk of complications.

In Shameer et al. [22], the authors assessed the performance of the algorithms using several key metrics. One of the primary indicators was accuracy, but additional metrics were also employed to give a more comprehensive view. Among these, precision and recall played a crucial role. These metrics are often combined to calculate the F1-score, which represents the harmonic mean of precision and recall, providing a balanced indicator of the model's performance. Another important metric was the area under the ROC curve (AUC-ROC), which helps evaluate the model's ability to distinguish between different classes. Finally, computational efficiency was measured through execution time and memory usage, which are critical when dealing with real-time data streams. These parameters allowed the authors to compare the scalability and practical applicability of the models in complex healthcare systems with large volumes of data.

3.2.6 RQ8: "What are the main research challenges specifically related to the generation and use of synthetic data in genomics?

This section analyzes the challenges related specifically to the use of SD in genomic research. While many issues in genomic data analysis are well-known, SD presents unique challenges that require further investigation. These include ensuring fidelity to real biological data, mitigating biases introduced by generative models, validating synthetic datasets effectively, and addressing ethical and regulatory concerns when using artificially generated patient-like data.

The study conducted by Chan et al. [5] highlights several challenges in inferring gene regulatory networks (GRNs) from single-cell data. These include the complexity and variability of single-cell data, technical noise such as dropout events, and the computational difficulty of handling large, high-dimensional datasets. Discretization of gene expression data and selecting accurate probability estimators also pose significant issues, impacting the reliability of network inference. Additionally, capturing nonlinear gene dependencies and interpreting network structures in context introduces further complexity, requiring advanced methods like PIDC to address these obstacles.

The study conducted by Song and Zhang [23] identifies several challenges in gene co-expression network analysis. A primary issue is the inefficiency of existing methods when applied to large genomic datasets, exacerbated by noisy and redundant gene pair similarities. Additionally, embedding techniques constrained by planarity often lead to false positives, complicating the identification of accurate gene interactions. Traditional clustering algorithms also fall short in recognizing multiscale modular structures within networks, limiting their ability to reveal biologically significant clusters across different scales. Moreover, improved methods are needed to effectively isolate compact, coherent gene clusters that reflect the complex hierarchical nature of biological systems, particularly in the context of diseases like cancer.

The study conducted by Li et al. [20], outlines several challenges in developing an effective cancer detection method. One major issue is the difficulty in identifying small amounts of tumor-derived cell-free DNA (cfDNA) amidst the abundance of normal cfDNA in plasma, particularly in early-stage cancers where tumor DNA levels are extremely low. Another challenge is the technical limitations of sequencing coverage,

as most public cfDNA data have low coverage, which reduces detection sensitivity. The pervasive nature of DNA methylation complicates the differentiation of cancer-specific signals from normal methylation patterns. Additionally, balancing the quality and quantity of methylation markers is crucial, as higher-quality markers may reduce the number of detectable tumor signals, particularly at low sequencing coverages. These challenges highlight the complexity of accurately detecting and quantifying tumor-derived cfDNA in a non-invasive and cost-effective manner.

In article conducted by El-Kebir [9], highlights several challenges in estimating tumor phylogenies from single-cell sequencing data. A major issue stems from the high levels of noise and sequencing errors, including dropouts, which complicate the accurate reconstruction of tumor evolution. Additionally, the computational demands of the task are considerable, as inferring phylogenies requires solving complex combinatorial problems that are both resource-intensive and time-consuming. Another significant challenge is balancing the rates of false positives and false negatives, as both can lead to inaccurate representations of tumor progression. Scalability is also a concern, as the computational load increases dramatically with larger datasets, which often contain numerous mutations and cells. Furthermore, the biological complexity of cancer, characterized by phenomena such as convergent evolution and the presence of multiple clones, further complicates the accurate modelling of tumor evolution. These challenges underscore the need for more advanced and efficient computational methods to enhance both the accuracy and scalability of tumor phylogeny inference from noisy single-cell data.

The study conducted by Zou et al. [32], identifies several challenges in utilizing deep learning for genomic research. One of the main issues is the requirement for large, carefully curated datasets to prevent biases that can distort model accuracy. Additionally, while deep learning models are highly effective, they are often more complex and less interpretable than traditional methods, making it harder to extract meaningful biological insights. Another challenge is the imbalance in genomic data, where disease-related variants are much rarer than neutral ones, complicating model evaluation. The significant computational resources required for training deep neural networks also limit accessibility. Finally, although deep learning excels in predictive accuracy, translating these predictions into actionable biological knowledge remains a major challenge.

The study conducted by Fisher and Mehta [11], outlines several key challenges in inferring species interactions from metagenomic data. One major difficulty is that correlations between species abundances do not necessarily indicate direct ecological interactions, leading to potential misinterpretation when using correlation-based methods. Moreover, metagenomic studies often provide only relative abundances, not absolute values, which complicates the estimation of interaction parameters in time-series models. Another challenge arises from experimental errors, such as sequencing inaccuracies and misclassification of reads into operational taxonomic units (OTUs), which can introduce bias. To address these issues, the authors developed the LIMITS algorithm, which applies sparse linear regression with bootstrap aggregation to reliably infer interaction networks despite these obstacles.

The study conducted by Yang and Michailidis [29] identifies significant difficulties in the integration of heterogeneous omics data for identifying biologically meaningful modules. One of the primary difficulties is the substantial heterogeneity among different omics data types—such as DNA methylation, gene expression, and microRNA

expression—which are generated through distinct experimental processes and contain varying degrees of noise. This variability complicates the extraction of coordinated biological signals from source-specific noise. Additionally, many existing integration methods depend heavily on prior biological knowledge, potentially introducing bias and limiting the discovery of novel insights. Another challenge lies in accurately capturing both shared and unique patterns across datasets while accounting for confounding factors that differ between sources. Addressing these issues requires advanced computational techniques capable of managing high-dimensional data and adapting to diverse levels of heterogeneity, as demonstrated by the non-negative matrix factorization method proposed in the study.

The study conducted by Wang et al. [27] identifies several challenges in selecting variables from high-dimensional genomic datasets. A primary issue arises from the presence of correlated and linearly dependent variables, which can cause traditional methods like the Lasso to exhibit instability and inconsistency in variable selection. In genomic datasets, where correlations between variables are common, standard approaches often select one variable from a group of correlated ones, potentially missing other relevant variables. Similarly, when variables are linearly dependent, traditional methods may prioritize combined variables rather than those with direct biological relevance. These challenges highlight the need for more robust methods that can effectively handle correlations and linear dependencies in the data. The Precision Lasso addresses these issues through advanced regularization techniques, offering a more stable and consistent approach to variable selection in genomic research.

In the study conducted by Haendel et al. [14], several significant challenges are highlighted in the development and implementation of the National COVID Cohort Collaborative (N3C). A key difficulty was the harmonization of electronic health record (EHR) data from numerous institutions, each employing different data models and formats, necessitating extensive efforts in standardization. Ensuring data privacy and security presented another critical challenge, as the large-scale integration of sensitive patient information required rigorous governance frameworks, regulatory compliance, and ethical oversight. Technical complexities also emerged in building a secure, centralized data enclave capable of managing vast datasets while facilitating collaborative research. Additionally, coordinating among multiple institutions and navigating legal agreements for data sharing posed logistical hurdles. These challenges were further exacerbated by the urgency of the COVID-19 pandemic, which demanded swift development and deployment.

The study conducted by Shameer et al. [22], highlights several significant challenges in the integration of real-time biomedical and healthcare data. A primary challenge lies in the aggregation and standardization of diverse data sources, including electronic medical records (EMRs), health-monitoring devices, and wellness data streams, each of which utilizes different formats and lacks uniform standards. Additionally, implementing real-time data capture while ensuring data quality, security, and privacy is a complex task, particularly when dealing with large-scale, heterogeneous datasets from varied populations. Another obstacle is the integration of patient-generated health data with clinical infrastructures that are typically not equipped to handle continuous, multi-modal data streams. Moreover, deriving actionable clinical insights from these vast datasets requires

sophisticated analytics and machine learning techniques capable of processing and interpreting real-time data for precision medicine applications.

4 Research area analysis and recent trends

Figure 11 shows the ten most globally cited documents between 2014 and 2024. These papers focus on SD generation in genomic cancer research, with keywords that highlight key research themes in this field. These keywords offer quick access to relevant literature and help us understand major scientific trends [26, 31].

A knowledge map of keyword co-occurrence was created to identify recurring themes in SD generation. Keywords such as "machine learning," "clustering," "deep learning," and "precision medicine" have gained prominence in recent years, reflecting a shift towards more complex, integrative data approaches and the use of AI in genomics.

The co-occurrence analysis highlighted a significant thematic shift, with early terms like "epistasis" and "proteomics" gradually giving way to "genomics," "multi-omics," and "precision medicine" (Fig. 12).

This trend suggests a growing focus on precision medicine and multi-omics integration, as researchers aim to develop advanced, predictive models for personalized cancer treatments.

The field of SD generation in genomics has evolved to include advanced methodologies, especially to address challenges like data scarcity and complexity in cancer research. Techniques such as *multi-omics integration* combine genomics, transcriptomics, and proteomics data to give a comprehensive view of cancer biology. This supports precision medicine by helping researchers create more accurate predictive models. *Machine learning and deep learning algorithms* play a pivotal role in synthetic genomic data research, enabling the generation, validation, and analysis of biologically meaningful datasets. Generative models such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) are widely employed to create synthetic genomic sequences that retain statistical properties of real patient data. Furthermore, deep learning techniques like Convolutional Neural Networks (CNNs) and Transformer-based architectures have demonstrated high accuracy in tumor classification and biomarker discovery, leveraging SD to enhance model robustness and generalization.

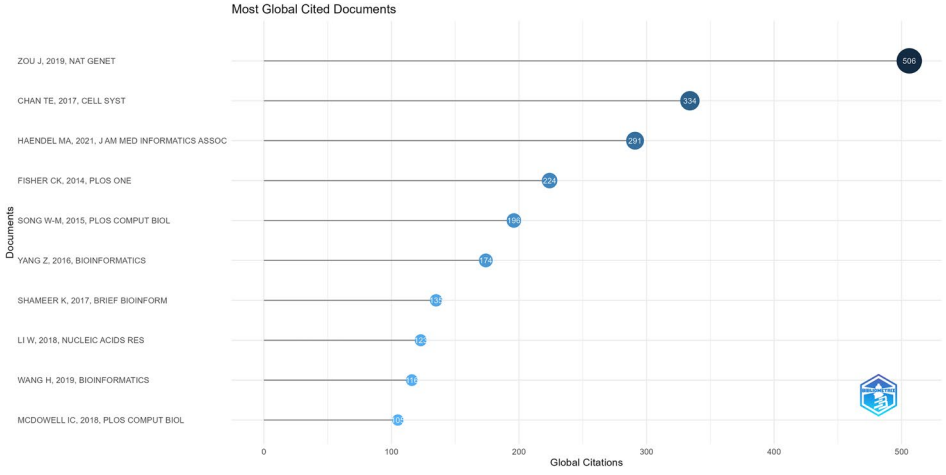

Fig. 11 Ten most globally cited documents from 2014 to 2024

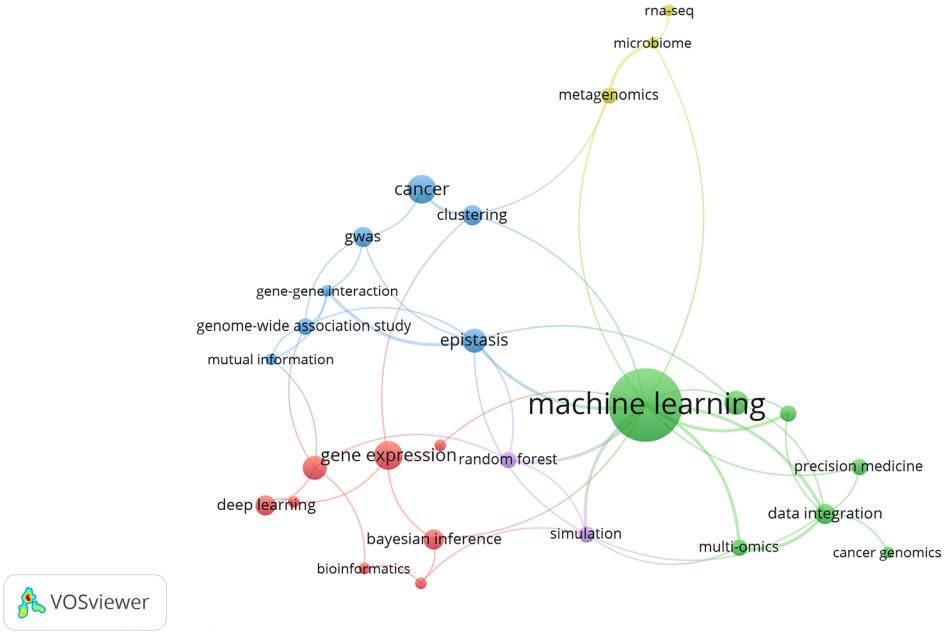

Fig. 12 High-frequency terms in author keywords for SD in genomics cancer publications (2014–2024)

Despite these advancements, there are still challenges in ensuring SDquality, particularly in controlling noise and achieving accurate simulations. To tackle these issues, models now use advanced regularization techniques, such as those in precision lasso, which effectively handle variable correlations and dependencies. These improvements are essential for the reliable application of SD in clinical settings.

The rise in open-access publications and international collaborations, especially among researchers in the United States, China, and the United Kingdom, has also been crucial for advancing SD research. Open access extends the reach of research findings, fostering global collaboration and accelerating innovation in precision medicine for cancer research.

This shift towards precision medicine and AI-driven models shows an increasing focus on developing highly tailored synthetic datasets, which enable targeted therapies and advance personalized treatment options.

5 Discussion and conclusion

This article explores how artificial intelligence is becoming a key player in generating SD for genomic oncology, a field where real clinical data is often hard to access. The analysis highlights specific algorithms, like deep neural networks for tumor classification, as well as advanced techniques for integrating genomic, transcriptomic, and proteomic data.

Some deep learning models, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and autoencoders, have proven to be transformative tools in genomic data analysis, offering powerful pattern-recognition abilities [1]. However, these models often require large, high-quality datasets and significant computational resources, which can limit their use in rare disease research or clinical settings with limited data [2].

The application of SD in the field of oncology to enhance the quality of clinical research and protect the confidentiality of patient data has attracted considerable interest within

Table 4 Summary of methods used in two or more studies

Method	Description	Number of papers
GANs (Generative adversarial networks)	Used for generating synthetic genomic data to train and validate models	12
Autoencoders	Applied for dimensionality reduction and feature extraction in genomic datasets	9
Random forest	Frequently used for feature selection and classification in multi-omics data	7
K-means clustering	Utilized to group biological samples based on synthetic data patterns	6
Bayesian networks	Applied for probabilistic modeling and causal inference in genomic studies	5
Support vector machines (SVM)	Used for classifying genomic data and detecting cancer subtypes	4
Lasso regression	Regularization technique employed in predictive modeling of gene expression	3
Principal component analysis (PCA)	Used for dimensionality reduction in high-dimensional genomic data	3

the medical community, Jacobs et al. [15]. One of the most exciting developments is the increased focus on personalizing SD to faithfully represent the genetic and molecular diversity across different types of tumors. This personalized approach could make it easier to create datasets specific to patient subgroups, allowing for the development of more targeted therapies. Moreover, as explainability techniques in deep learning are being more widely adopted, interpreting results is becoming easier—a crucial factor when applying this data in clinical settings [28].

As genomic and clinical datasets continue to expand, these advanced algorithms will play a vital role in gaining insights from data, improving patient outcomes, and turning complex information into actionable knowledge for precision medicine [13]. Our findings confirm that SD generation algorithms can effectively replicate meaningful genomic patterns, as demonstrated by high AUC and precision-recall scores. More importantly, their application in research settings has shown promising results, particularly in biomarker discovery and model training for multi-omics analysis. However, while these techniques offer significant advantages, challenges such as model interpretability, integration into clinical pipelines, and long-term validation remain open issues that require further investigation. By revealing hidden patterns in high-dimensional data, these models can aid in patient stratification, though further work is needed to ensure they adapt well to new and constantly evolving patient data in real-time clinical settings.

To provide a clearer perspective on the most frequently used computational approaches in SD research, we identified methods that appeared in at least two studies. Table 4 summarizes these methods, emphasizing their role in genomic data analysis and SD validation.

This summary highlights the predominance of deep learning models such as GANs and autoencoders in SD generation, as well as the widespread use of machine learning techniques for feature selection, classification, and dimensionality reduction.

One of the key challenges in SD generation is ensuring that the generated datasets maintain biological fidelity while minimizing noise and artifacts introduced by generative models. The accuracy of SD depends on how well it replicates real genomic variability while avoiding biases that could compromise downstream analysis.

Another important issue is the ethical and regulatory aspect of synthetic data usage. Although synthetic datasets can help mitigate privacy concerns by eliminating direct patient identifiers, regulatory agencies require rigorous validation and transparency to ensure that these datasets do not introduce misleading biases in clinical applications.

Current trends show a growing level of specialization, with a strong focus on developing models tailored to specific tumor subtypes and the search for complex biomarkers.

Another crucial aspect emerging from our analysis is the adaptability of different algorithms to various genomic data types and the importance of hybrid approaches in SD research. While deep learning models such as GANs and VAEs excel in generating synthetic genomic data, their effectiveness is highly dependent on the availability of large, well-balanced datasets. In contrast, tree-based methods and probabilistic models, such as Random Forest and Bayesian Networks, offer greater interpretability but struggle with high-dimensional omics data. Similarly, graph-based methods like PIDC and MEGENA play a critical role in gene network inference but require complex preprocessing and parameter tuning.

Our findings indicate that no single algorithm provides optimal performance across all genomic contexts. Instead, hybrid approaches combining deep learning and probabilistic inference have shown promising results, particularly in tumor evolution modeling. Ensemble learning strategies have also emerged as a powerful tool to improve biomarker discovery, while feature selection techniques such as Precision Lasso have proven effective in reducing dimensionality and optimizing SD quality. These insights highlight the need for flexible, multi-method strategies that can dynamically adapt to different genomic data challenges, ensuring both accuracy and biological interpretability.

Looking to the future, it will be essential to create even more advanced algorithms capable of capturing genomic complexities and making SD a practical and indispensable tool for precision medicine.

Author contributions

Maria Frasca, Valentina De Nicoló, and Agnese Graziosi contributed to the study design, data analysis, and manuscript writing. Gianluca Gazzaniga, Davide La Torre, and Arianna Pani contributed significantly to the critical revision of the scientific and methodological content.

Funding

The authors declare that this study did not receive any specific funding from public, commercial, or non-profit funding agencies.

Data availability

The data on which this analysis is based were collected from the Scopus database.

Declarations

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare no conflict of interest related to this study.

Received: 12 November 2024 / Accepted: 18 June 2025

Published online: 15 July 2025

References

1. Alharbi WS, Rashid M. A review of deep learning applications in human genomics using next-generation sequencing data. Hum Genomics. 2022;16(1):26.
2. Alzubaidi L, Bai J, Al-Sabaawi A, Santamaría J, Albahri AS, Al-dabbagh BSN, Fadhel MA, Manoufali M, Zhang J, Al-Timemy AH, Duan Y, Abdullah A, Farhan L, Lu Y, Gupta A, Albu F, Abbosh A, Gu Y. A survey on deep learning tools dealing with data

scarcity: definitions, challenges, solutions, tips, and applications. J Big Data. 2023. https://doi.org/10.1186/s40537-023-00727-2.
3. Ballew BS. Elsevier's Scopus® database. J Electron Resour Med Libr. 2009;6(3):245–52.
4. Chan D, Shao X, Dumargne M-C, Aarabi M, Simon M-M, Kwan T, Bailey JL, Robaire B, Kimmins S, Gabriel MCS, et al. Customized methylc-capture sequencing to evaluate variation in the human sperm DNA methylome representative of altered folate metabolism. Environ Health Persp. 2019;127(8):087002.
5. Chan TE, Stumpf MPH, Babtie AC. Gene regulatory network inference from single-cell data using multivariate information measures. Cell Syst. 2017;5(3):251–67.
6. Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. Nat Biomed Eng. 2021;5(6):493–7.
7. Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J. Generating multi-label discrete patient records using generative adversarial networks. In: Machine learning for healthcare conference, 2017. p. 286–305. PMLR.
8. Das S, Mazumder S, Alam N, Vernekar M, Dam A, Bhowmick AK, Hajra S, Das JK, Basu B. Precision oncology in the era of genomics and artificial intelligence. J Curr Oncol Trends. 2024;1(1):22–30.
9. El-Kebir M. SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error. Bioinformatics. 2018;34(17):i671–9.
10. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nat Rev Genet. 2014;15(6):409–21.
11. Fisher CK, Mehta P. Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. PLoS One. 2014;9(7):e102451.
12. Frasca M, La Torre D, Repetto M, De Nicolò V, Pravettoni G, Cutica I. Artificial intelligence applications to genomic data in cancer research: a review of recent trends and emerging areas. Discov Anal. 2024;2:10. https://doi.org/10.1007/s44257-024-00017-y.
13. Ghebrehiwet I, Zaki N, Damseh R, Mohamad MS. Revolutionizing personalized medicine with generative AI: a systematic review. Artif Intell Rev. 2024;57(5):128.
14. Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, Payne PRO, Pfaff ER, Robinson PN, Saltz JH, et al. The national Covid cohort collaborative (N3C): rationale, design, infrastructure, and deployment. J Am Med Informatics Assoc. 2021;28(3):427–43.
15. Jacobs F, D'Amico S, Benvenuti C, Gaudio M, Saltalamacchia G, Miggiano C, De Sanctis R, Porta MGD, Santoro A, Zambelli A. Opportunities and challenges of synthetic data generation in oncology. JCO Clin Cancer Inform. 2023;7:e2300045.
16. Jiang Y, Mosquera L, Jiang B, Kong L, El Emam K. Measuring re-identification risk using a synthetic estimator to enable data sharing. PLOS One. 2022;17(6):e0269097. https://doi.org/10.1371/journal.pone.0269097.
17. Kokosi T, Harron K. Synthetic data in medical research. BMJ Med. 2022;1(1):e000167. https://doi.org/10.1136/bmjmed-2022-000167.
18. Leung ML, Davis A, Gao R, Casasent A, Wang Y, Sei E, Vilar E, Maru D, Kopetz S, Navin NE. Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer. Genome Res. 2017;27(8):1287–99.
19. Li W, Kim M, Zhang K, Chen H, Jiang X, Harmanci A. COLLAGENE enables privacy-aware federated and collaborative genomic data analysis. Genome Biol. 2023;24(1):204.
20. Li W, Li Q, Kang S, Same M, Zhou Y, Sun C, Liu C-C, Matsuoka L, Sher L, Wong WH, et al. Cancerdetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free dna methylation sequencing data. Nucl Acids Res. 2018;46(15):e89–e89.
21. Patki N, Wedge R, Veeramachaneni K. The synthetic data vault. In: 2016 IEEE international conference on data science and advanced analytics (DSAA). p. 399–410. IEEE 2016.
22. Shameer K, Badgeley MA, Miotto R, Glicksberg BS, Morgan JW, Dudley JT. Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams. Brief Bioinform. 2017;18(1):105–24.
23. Song W-M, Zhang B. Multiscale embedded gene co-expression network analysis. PLoS Comput Biol. 2015;11(11):e1004574.
24. Sweileh WM. Research trends on human trafficking: a bibliometric analysis using Scopus database. Global Health. 2018;14:1–12.
25. Torkamani A, Andersen KG, Steinhubl SR, Topol EJ. High-definition medicine. Cell. 2017;170(5):828–43. https://doi.org/10.1016/j.cell.2017.08.007.
26. Vargas-Quesada B, Chinchilla-Rodríguez Z, Rodriguez N. Identification and visualization of the intellectual structure in graphene research. Front Res Metrics Anal. 2017;2:7.
27. Wang H, Lengerich BJ, Aragam B, Xing EP. Precision lasso: accounting for correlations and linear dependencies in high-dimensional genomic data. Bioinformatics. 2019;35(7):1181–7.
28. Wysocka M, Wysocki O, Zufferey M, Landers D, Freitas A. A systematic review of biologically-informed deep learning models for cancer: fundamental trends for encoding and interpreting oncology data. BMC Bioinform. 2023;24(1):198.
29. Yang Z, Michailidis G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. Bioinformatics. 2016;32(1):1–8.
30. Yelmen B, Decelle A, Ongaro L, Marnetto D, Tallec C, Montinaro F, Furtlehner C, Pagani L, Jay F. Creating artificial human genomes using generative neural networks. PLOS Genet. 2021;17(2):e1009303. https://doi.org/10.1371/journal.pgen.1009303.
31. Zhang J, Qi Y, Zheng F, Long C, Zuxun L, Duan Z. Comparing keywords plus of wos and author keywords: a case study of patient adherence research. J Assoc Inf Sci Technol. 2016;67(4):967–72.
32. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti Amalio. A primer on deep learning in genomics. Nat Genet. 2019;51(1):12–8.

Publisher's Note