Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Landscape and Emerging Directions of Single-Cell Genetic Variation Research Revealed by LLM-Assisted Analysis

Hechi Zou
Tianjin Medical University
jingjing Zhang
Tianjin Medical University
Jiawen Mao
Tianjin Medical University
Nianling Yang
Tianjin Medical University
Yongjian Wei
Tianjin Medical University
Xiaoran Bai
Tianjin Medical University
Hui Zhang
Tianjin Medical University
Weiwei Gao
Hunan Weishi Biotechnology Research Institute
Xianfu Yi
yixianfu@tmu.edu.cn

Tianjin Medical University

Research Article

Keywords: Large language models (LLMs), Single-cell genetic variation, Single-cell sequencing, Tumor heterogeneity, Multi-omics integration

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11657247/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Landscape and Emerging Directions of Single-Cell Genetic Variation Research Revealed by LLM-Assisted Analysis

Hechi Zou[1,#], Jingjing Zhang[1,#] , Jiawen Mao[1] ,Nianling Yang[1], Yongjian Wei[1] , Xiaoran Bai[2], Hui Zhang[3], Weiwei Gao[4,*], Xianfu Yi[1,*]

[1]Department of Bioinformatics, School of Basic Medical Sciences, Tianjin Medical University, Tianjin 300070, China

[2]Department of Biopharmaceuticals, School of Pharmacy, Tianjin Medical University, Tianjin 300070, China

[3]Department of Clinical Medicine, The Second School of Clinical Medicine, Tianjin Medical University, Tianjin 300070, China

[4]Hunan Weishi Biotechnology Research Institute, Changsha 410000, China

[#]Hechi Zou, Jingjing Zhang have contributed equally to this research

*** Correspondence:**

Weiwei Gao, gww@w-ecology.com; Xianfu Yi, yixianfu@tmu.edu.cn

Abstract

Background Single-cell genetic variation research has expanded rapidly with advances in sequencing technologies and computational approaches and is increasingly important for investigating cellular heterogeneity, clonal evolution, and disease-associated genomic alterations. However, key methodological and experimental information is often dispersed across full-text publications, limiting systematic characterization of the field. To address this, we combined large-scale literature analysis with full-text structured information extraction assisted by the large language model (LLM) ChatGPT-4o to characterize technological, methodological, and application trends in single-cell genetic variation research.

Results Information on sequencing technologies, variant types, disease applications, species, and variant detection tools was systematically extracted and organized. ChatGPT-4o-assisted extraction achieved a mean correctness of 82.50% and a mean completeness of 83.05% compared with manually curated annotations. Single-cell genetic variation research has expanded rapidly, accompanied by increasing international and institutional collaboration. scRNA-seq remains the predominant technology, while scATAC-seq, scDNA-seq, spatial transcriptomics, and multi-omics combinations are increasingly adopted. Cancer is the major application area, with tumor microenvironment, prognosis, and immunotherapy as prominent themes. SNVs remain the most extensively studied variant type, whereas combined SNV-CNV analyses are becoming more common, particularly in cancer studies. Bulk-oriented tools such as GATK remain widely used, while single-cell-specific methods are increasingly adopted to address low coverage, allelic dropout, and technical noise.

Conclusions Single-cell genetic variation research is evolving toward integrated multi-omics, multi-variant analysis, and more specialized single-cell computational strategies. The increasing adoption of multi-omics technologies and single-cell-specific methods reflects the growing complexity and maturity of the field. Combining large-scale literature analysis with

LLM-assisted full-text information extraction provides a systematic approach for capturing complex methodological and biological information and identifying emerging directions in single-cell genomics.

Keywords

Large language models (LLMs); Single-cell genetic variation; Single-cell sequencing; Tumor heterogeneity; Multi-omics integration

Introduction

Genetic variation refers to alterations in genomic sequences relative to a reference genome and represents a fundamental driver of biological evolution, tissue development, and disease progression[1]. With the rapid advancement of next-generation sequencing (NGS) technologies, whole-genome sequencing (WGS) has enabled the identification of a large number of rare variants and provided essential foundations for characterizing genome architecture and exploring the functional potential of previously unexplored genomic regions[2, 3]. However, conventional WGS approaches are typically performed on bulk samples containing large numbers of cells and primarily capture averaged genetic signals at the population level, limiting their ability to resolve cell-to-cell genetic heterogeneity and detect low-frequency mutations. Since genetic differences among individual cells represent a critical basis for tissue development, disease progression, and tumor clonal evolution[4], deciphering genetic variation at the single-cell level has become increasingly important in the context of precision medicine.

With the rapid development of single-cell sequencing technologies and the continuous accumulation of single-cell datasets, researchers have been able to characterize cellular heterogeneity, rare cell populations, and tumor microenvironmental features at unprecedented resolution[5, 6]. Furthermore, genetic variation information can be extracted directly from single-cell sequencing data, enabling the identification of low-frequency somatic mutations that are difficult to detect using conventional bulk sequencing approaches, the characterization of genetic differences among individual cells, and the inference of clonal evolution processes[7]. These advances provide a higher-resolution perspective for understanding complex disease mechanisms, tumor evolution, and tissue development[8]. In addition, increasing efforts have been devoted to developing specialized computational tools for single-cell genetic variation detection based on different data modalities and computational strategies, including Cellsnp-lite[9], Souporcell[10], SComatic[11], and Monopogen[12].Numerous studies have investigated the roles of single-cell genetic variation in disease initiation and progression, demonstrating its critical value in deciphering cellular heterogeneity, tumor clonal evolution, and cell lineage tracing [13–15]. However, single-cell genetic variation research is highly complex and heterogeneous, involving diverse single-cell sequencing technologies, multiple types of genetic variants, various analytical tools, and a wide range of disease contexts[16]. The relevant information is often dispersed across experimental designs, analytical workflows, and result interpretations within full-text publications, making it difficult to comprehensively capture such information solely through bibliographic records or keyword-based analyses. Moreover, with the rapid expansion of the literature, a systematic understanding of the overall

developmental trajectory, research hotspots, technological evolution, and future directions of single-cell genetic variation research remains lacking. Therefore, beyond conventional bibliometric approaches, there is an urgent need to incorporate methods capable of processing unstructured textual information to achieve a more comprehensive and in-depth characterization of this emerging field[17].

Large language models (LLMs) have demonstrated advantages in processing complex textual information and extracting structured knowledge, enabling the identification and organization of dispersed key information from full-text publications[18]. Based on this capability, this study integrates conventional bibliographic analysis with ChatGPT-4o-assisted full-text information extraction. Specifically, we systematically characterize the overall development trends, collaboration patterns, keyword evolution, and scientific impact of single-cell genetic variation research through bibliographic analysis, while simultaneously extracting and structuring key information from full texts, including sequencing technologies, variant types, disease categories, and species origins. This integrated strategy enables a more comprehensive understanding of the current landscape, emerging hotspots, and future directions of single-cell genetic variation research from both macro-level trends and detailed research characteristics[19, 20].

Materials and methods

Data Collection and preprocessing

A comprehensive literature search was performed using the Web of Science Core Collection database via the ISI Web of Knowledge platform. Publications from January 1, 1998 to December 31, 2025 were retrieved using a predefined search strategy focusing on studies related to single-cell genetic variation. The search query was as follows: (((variant OR variation OR "copy number variation" OR mutation OR polymorphism OR "genetic alteration" OR "structural variation" OR deletion OR duplication OR inversion OR insertion) AND (genetic OR genomic OR gene OR chromosome)) AND ("single cell" OR "single cell sequencing" OR "single cell genomics" OR "single nucleus sequencing")). This search initially identified 11,269 publications from the Web of Science Core Collection. After removing duplicate records based on unique Web of Science accession numbers (UT), 11,268 unique publications were retained for subsequent analyses. All downstream statistical analyses and data visualizations were performed using R software (version 4.6.0).

Analysis of Annual Publication Trends Using Log-Transformed Linear Regression Models

Based on the document type (Document Type), publications were classified into three categories: Article, Review, and Others. Document types other than Article and Review were grouped into the Others category. The annual publication counts for each document category were calculated to evaluate temporal changes in publication patterns within the field of single-cell genetic variation research. To assess annual growth trends, the annual publication counts were transformed using natural logarithmic transformation after adding one to each count to avoid undefined values caused by zero counts. Ordinary least squares (OLS) linear regression was then applied to model the temporal trends according to the following equation:

$$\log(y_i + 1) = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where y_i represents the annual publication count of a specific document type in year i, $X_i$represents the corresponding publication year, $\beta_0$is the intercept, $\beta_1$represents the annual temporal trend coefficient, and ε_i denotes the random error term. Model parameters were estimated using the ordinary least squares method by minimizing the residual sum of squares. The goodness-of-fit of the regression model was evaluated using the coefficient of determination R^2.

Analysis of Publication Characteristics and Research Trends

Author and country contribution analysis

Author information was extracted from the "Author Full Names" field in the Web of Science records. The number of authors per publication was calculated, and the annual average number of authors was summarized to evaluate temporal changes in research team size. Country/region information was extracted from the "Reprint Addresses" field, representing the affiliation country/region of the corresponding author. Annual publication output and proportion were calculated for each country/region to characterize the geographic distribution of research contributions.

Citation impact analysis

Citation impact analysis was performed based on the "Times Cited, All Databases" field. The total citation count of each publication was calculated and log10-transformed using:

$$Citation_i^{log} = \log_{10}(Times\ Cited_i + 1)$$

to reduce the influence of highly skewed citation distributions during visualization. Citation distributions among different document types were compared based on the "Document Type" field. In addition, total citations, publication output, and annual citation proportions were summarized according to corresponding author countries/regions.

To investigate the association between citation impact and research team size, publications were ranked according to total citation counts and categorized into citation percentile groups (0–10%, 10–20%, ..., 90–100%). Representative citation groups, including Top 1%, Middle 50–51%, and Bottom 1%, were further selected. The number of authors per publication was calculated based on the "Author Full Names" field, and differences in author numbers among citation groups were evaluated using the Wilcoxon rank-sum test.

Journal distribution and quartile analysis

Journal quartiles were assigned according to the 2024 Journal Citation Reports (JCR) by matching standardized ISSN/eISSN identifiers from Web of Science records. Journals were classified as Q1–Q4, while unmatched journals were grouped as “Others.” Annual publication counts and proportions by quartile were calculated, and the top 10 journals by publication output were analyzed to assess temporal changes in major publication venues.

Collaboration network analysis

Collaboration networks were constructed using the Web of Science “Addresses” field.

Country/region and institutional information were extracted from author affiliations, and collaboration patterns were summarized according to participating countries or institutions. The top 20 countries/regions and top 10 institutions by publication output were used to construct undirected weighted networks, where nodes represented countries/regions or institutions and edge weights represented the number of collaborative publications.

Keyword co-occurrence and country-specific research topic analysis

Author Keywords were cleaned, standardized, and used to calculate keyword frequencies and construct co-occurrence networks. Nodes represented keywords, node size reflected frequency, and edge weights indicated the number of co-occurrences within the same publication, allowing major research themes and their relationships to be characterized.

Keyword analysis was also performed for highly productive authors to identify their major research topics. In addition, the top five countries by publication output were compared using country-specific keyword frequencies after removing duplicate keywords within each publication, enabling assessment of similarities and differences in national research focuses.

Large language model-assisted extraction of structured information from full texts

After identifying the included publications, all available full-text PDF files were downloaded and organized for structured information extraction using a large language model (LLM). Full-text PDFs were first converted into machine-readable text, followed by information extraction through the OpenAI API using ChatGPT-4o. The model was configured with a temperature of 0 and was not further fine-tuned. We extracted key information from the full text, including sequencing technologies, variant types, disease categories, and species origins. The extracted information was formatted according to a predefined JSON schema and subsequently parsed into structured tables for downstream analysis.

Reliability assessment of LLM-assisted structured information extraction

To evaluate the reliability of LLM-based structured information extraction, a manually curated validation set was established. Thirty publications were randomly selected from the included literature and independently reviewed by three researchers. For each publication, manually identified items were compared with the corresponding ChatGPT-4o extraction results for four information categories: sequencing technology, variant type, disease category, and species. An item was considered part of the manual consensus when it was independently identified by at least two of the three reviewers.

Extraction performance was evaluated in terms of correctness and completeness. Correctness was assessed at the publication–category level. A ChatGPT-4o extraction was considered correct only when all extracted items were supported by the manual consensus; if any extracted item was not supported by at least two reviewers, the extraction was scored as incorrect. Correctness was therefore assigned a binary score of 1 or 0 for each publication and information category, and the mean correctness was calculated across the 30 publications. Completeness was calculated as the proportion of manually identified consensus items that were successfully captured by ChatGPT-4o. For each publication and information category, completeness was defined as the number of consensus items also identified by ChatGPT-4o divided by the total number of manually identified consensus items. Mean completeness was

subsequently calculated across the 30 publications for each information category.

LLM-assisted characterization of single-cell genetic variation technologies, applications, and research evolution

Based on the structured information extracted by the large language model, sequencing technologies, disease applications, genetic variation types, and variant detection strategies were systematically characterized in single-cell genetic variation research.

Sequencing technologies were classified into five categories: Single-cell, Bulk DNA, Bulk RNA, Bulk Chromatin, and Spatial transcriptomics. Temporal trends, technology combinations, sequencing platforms, and country-specific patterns were analyzed, with further evaluation of major single-cell approaches, including scRNA-seq, scATAC-seq, and scDNA-seq. Disease-related information was integrated to characterize the application of single-cell sequencing across different disease categories and research topics.

Genetic variation patterns were analyzed based on four major variant types: SNV, CNV, SV, and Indels. Their distributions, temporal trends, and co-occurrence relationships were evaluated. Variant detection tools were classified into single-cell and bulk tools, and their usage patterns were compared. Finally, sequencing technologies, variant types, detection tools, and disease categories were integrated to characterize the research landscape and evolutionary trends of single-cell genetic variation research.

Results

Single-cell genetic variation research is rapidly expanding, with an international collaboration landscape increasingly centered on China and the USA.

We first classified the retrieved publications related to single-cell variation research into three document types: Article, Review, and Others. Articles represented the dominant publication type, accounting for 85.9% of all publications (Fig. S1A), indicating that research output in this field is primarily driven by original studies. The emergence of single-cell transcriptomics in 2009[21]provided an important foundation for subsequent single-cell variation research. Annual publication trends showed that Article-type publications increased markedly after 2016, rising from 106 publications in 2009 to 1,786 in 2025 and exhibiting a significant exponential growth pattern ($R^2 = 0.90$, $p < 0.001$; Fig. 1A; Table S1). This trend reflects the rapid expansion of single-cell variation research alongside advances in single-cell sequencing technologies.

Meanwhile, the number of authors per publication gradually increased over time. During 1998–2009, the median number of authors remained mainly between 4 and 5, whereas it increased to 5–9 during 2010–2025 (Fig. 1B; Table S2). This trend suggests a shift from relatively small research teams toward broader collaborative efforts involving larger numbers of researchers.

Temporal analysis of contributions from the top 10 countries/regions ranked by publication output showed that the USA dominated the early development of the field. Among

these countries/regions, the USA accounted for 53.0% of publications in 2009, followed by a gradual decline to 24.0% in 2025. In contrast, China contributed only 4.2% in 2009 but increased rapidly to 60.0% in 2025. These findings indicate a shift from early USA dominance toward a more geographically distributed international research landscape, with China emerging as an increasingly important contributor (Fig. 1C).

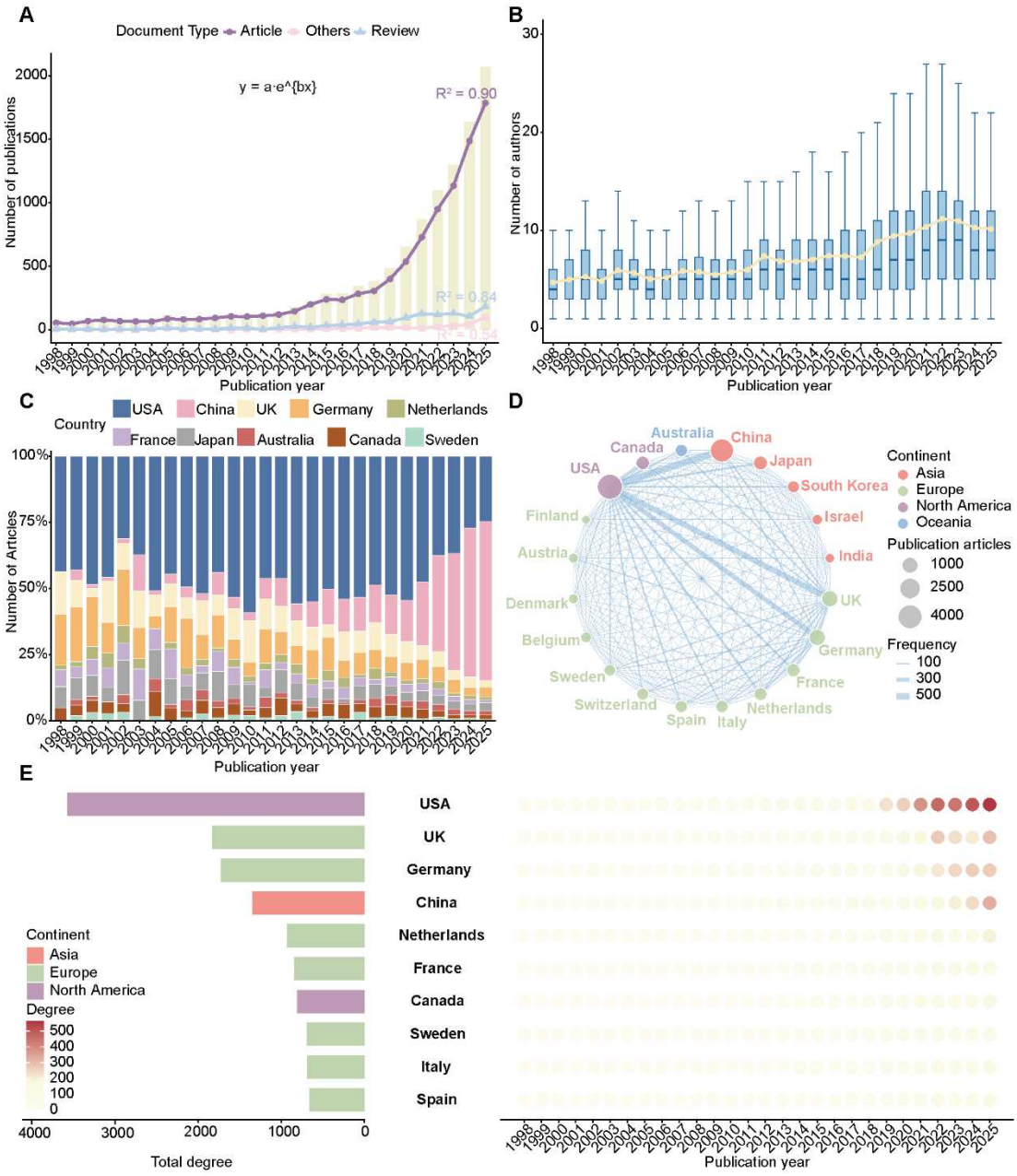


Fig. 1 Publication Growth and International Collaboration in Single-Cell Genetic Variation Research. **A** Annual publication trends across different document types. **B** Temporal changes in the number of authors per publication. **C** Annual publication contributions of major countries/regions. **D** International collaboration network among major contributing countries/regions. **E** Temporal evolution of the top 10 countries/regions ranked by Degree

centrality in the international collaboration network.

International collaboration increased over time, with the proportion of internationally collaborative publications rising from 4.7%–22.0% before 2009 to 23.1%–34.2% after 2009 (Fig. S1B; Table S3). The strongest country-level collaboration was observed between China and the USA (weight = 513; Fig. 1D). Network centrality also increased substantially, with the degree of the USA rising from 3–26 before 2009 to 22–561 after 2009, and that of China from 0–5 to 1–330 (Fig. 1E). Internationally collaborative publications also showed higher citation impact than non-collaborative publications (median citations: 18 vs. 10; $p < 0.001$), suggesting that cross-national collaboration may enhance research visibility and impact (Fig. S1C).

At the institutional level, highly productive institutions were mainly concentrated in the USA and China, with Harvard University (14.2%) and Stanford University (9.6%) among the leading institutions (Fig. S1D). Broad Institute and Harvard University showed relatively high citation impact (median citations = 41 and 40, respectively; Fig. S1E). Strong institutional collaborations were observed between Harvard University and Massachusetts Institute of Technology (weight = 337) and between Broad Institute and Harvard University (weight = 306), highlighting their central roles in research collaboration and knowledge exchange (Fig. S1F).

Academic impact of single-cell variation research is increasingly associated with larger collaborative teams and high-impact journals

To characterize the academic impact of single-cell genetic variation research, we first compared citation patterns across document types. Review articles showed higher citation impact than Articles and Others, with median citation counts of 19, 11, and 3, respectively (Fig. S2A), consistent with the generally higher citation frequency of review articles due to their broader synthesis of existing knowledge[22].

We next examined the temporal evolution of citation influence among the top 10 countries ranked by total citations. The USA dominated during the early stage, contributing 37.1%–80.0% of citations before 2009 and 33.62%–73.09% thereafter. In contrast, China increased from 0.09%–3.49% before 2009 to 1.02%–49.19% after 2009, indicating a substantial rise in citation influence consistent with its growing publication output (Fig. 2A).

Highly cited publications were also associated with larger research teams. Top 1% publications had significantly more authors than Middle 50–51% and Bottom 1% publications (median: 11 vs. 7 vs. 7; $p < 0.01$; Fig. 2B). Similarly, publications within the top 20% citation percentile had higher average author numbers than lower-ranked publications (8 vs. 7; Fig. S2B), suggesting that greater academic impact is often associated with broader collaboration. Among Top 1% highly cited publications, Science & Technology was the most prevalent research area (30.0%), highlighting the importance of technological innovation and interdisciplinary integration (Fig. 2C).

Journal distribution analysis showed that 65.0% of publications appeared in JCR Q1 journals (Fig. 2D), with the proportion increasing from 38.8%–56.6% before 2009 to 45.5%–73.0% after 2009 (Fig. S2C). Publications in Q1 journals also involved more authors than those

in non-Q1 journals (median: 8 vs. 6; $p < 0.01$; Fig. S2D). The top 10 journals were predominantly JCR Q1 titles, including *Nature Communications* and *Nature*. The contribution of *Nature Communications* increased from 0% in 2009 to 17.9% in 2025, while *Frontiers in Immunology* increased to 26.9%, further reflecting the growing presence of single-cell genetic variation research in high-impact journals (Fig. 2E).

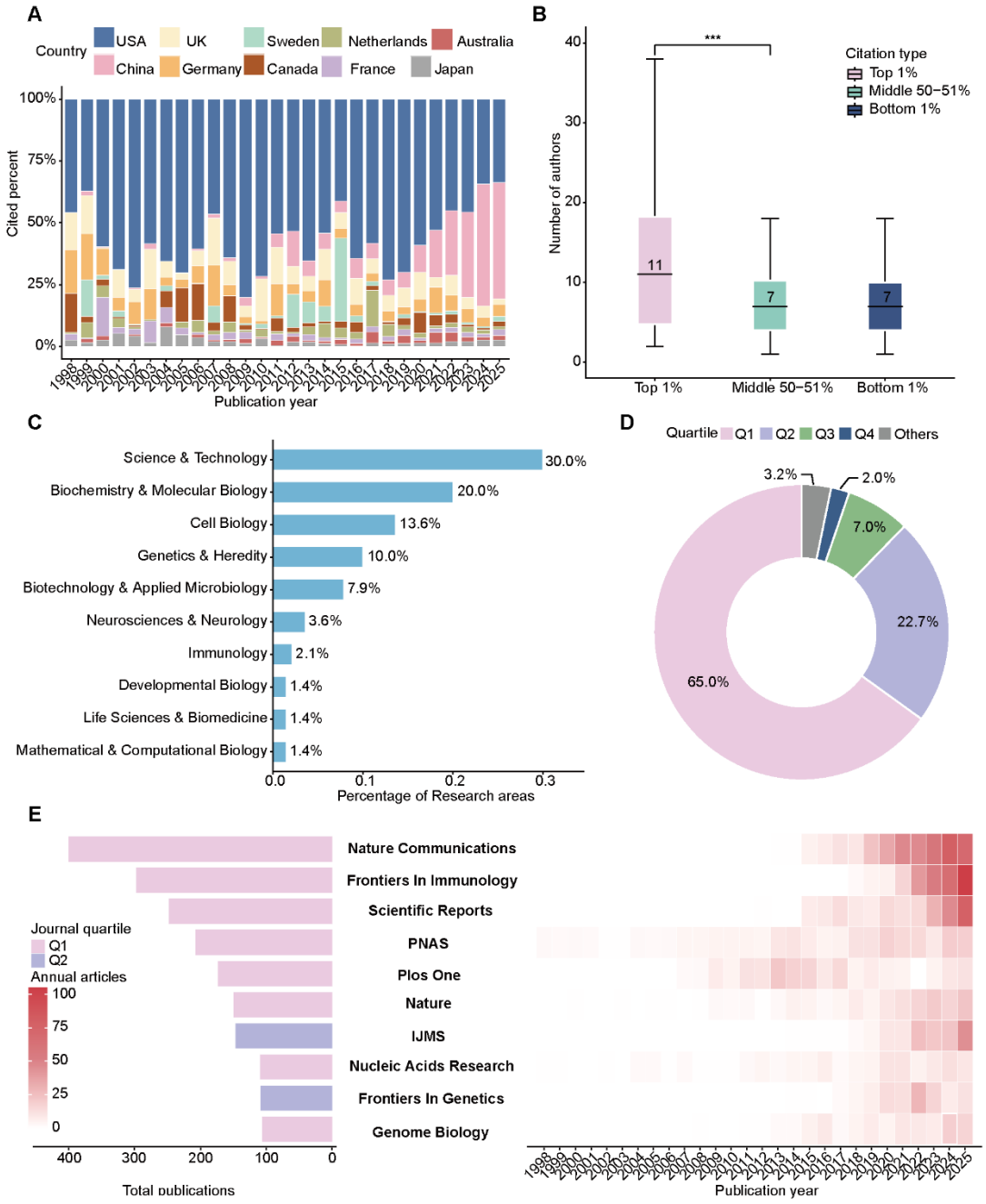


Fig. 2 Citation Impact and Journal Distribution in Single-Cell Genetic Variation Research. **A** Temporal changes in the proportion of annual citations contributed by the top 10 countries/regions ranked by total citation counts. **B** Distribution of author counts across different citation impact groups. *** indicates a statistically significant difference at $p < 0.01$. **C** Top 10 research areas represented among Top 1% highly cited publications. **D** Distribution

of JCR quartile categories among journals publishing single-cell genetic variation studies. **E** Top 10 journals ranked by publication output and their annual publication trends.

Single-cell variation research is increasingly driven by technological advances and has formed a scRNA-seq-centered research landscape

To characterize major research hotspots, we first identified the Top 10 highly cited publications. Representative studies included the 2015 *Science* article "Tissue-based map of the Human Proteome" (13,165 citations) and the 2018 *Nature Biotechnology* article "Integrating single-cell transcriptomic data across different conditions, technologies, and species" (9,228 citations) (Fig. S3A; Table S4). These highly cited studies focused on large-scale omics resources, single-cell data integration, and cross-condition and cross-technology analyses, underscoring the importance of technological advances and data integration in the field.

We next analyzed the top 10 most productive authors, who were mainly concentrated in the USA and China, accounting for 50.0% and 30.0% of the top 10 authors, respectively (Fig. S3B; Table S5). Aviv Regev from Harvard University ranked first with 31 publications. Her major research areas were Science & Technology, Cell Biology, and Biochemistry & Molecular Biology, accounting for 41.7%, 22.2%, and 16.7%, respectively (Fig. S4A). Keyword analysis showed that her research focused on single-cell omics, cell atlas construction, and cellular state characterization, with recent expansion toward spatial omics (Fig. S4B). As a founding co-chair of the Human Cell Atlas initiative, Regev has made major contributions to cell atlas construction and the characterization of cellular heterogeneity[23]. Fuchou Tang from Peking University published 26 articles, with Science & Technology, Cell Biology, and Genetics & Heredity accounting for 26.9%, 23.1%, and 15.4% of his research areas, respectively (Fig. S4C). His research mainly focused on tumor heterogeneity, colorectal cancer, and early embryonic development using single-cell sequencing technologies (Fig. S4D).

We extracted Author Keywords and identified the Top 30 most frequent terms, among which scRNA-seq was the most common, accounting for 20.1% of occurrences (Fig. 3A). Most high-frequency keywords increased over time, particularly scRNA-seq, tumor microenvironment, and single-cell sequencing, which rose from 0 occurrences in 2009 to 357, 130, and 54 in 2025, respectively (Fig. 3B). These trends indicate a shift from technology-centered studies toward broader biological questions involving tumor microenvironment, cellular heterogeneity, and multi-omics integration.

Keyword co-occurrence analysis showed that scRNA-seq most frequently co-occurred with tumor microenvironment (92 occurrences), followed by RNA-seq (48) and machine learning (37), highlighting the close links between single-cell transcriptomics, disease research, and computational analysis (Fig. 3C). Among the three most productive countries, 9 of the Top 15 keywords were shared, including scRNA-seq, prognosis, tumor microenvironment, and single-cell, indicating broadly similar research interests (Fig. 3D). Nevertheless, distinct national focuses were observed: China showed greater emphasis on lung adenocarcinoma and clear cell renal cell carcinoma, Germany on colorectal and breast cancer, and the USA on genetics-related topics such as Mendelian randomization (Fig. 3E).

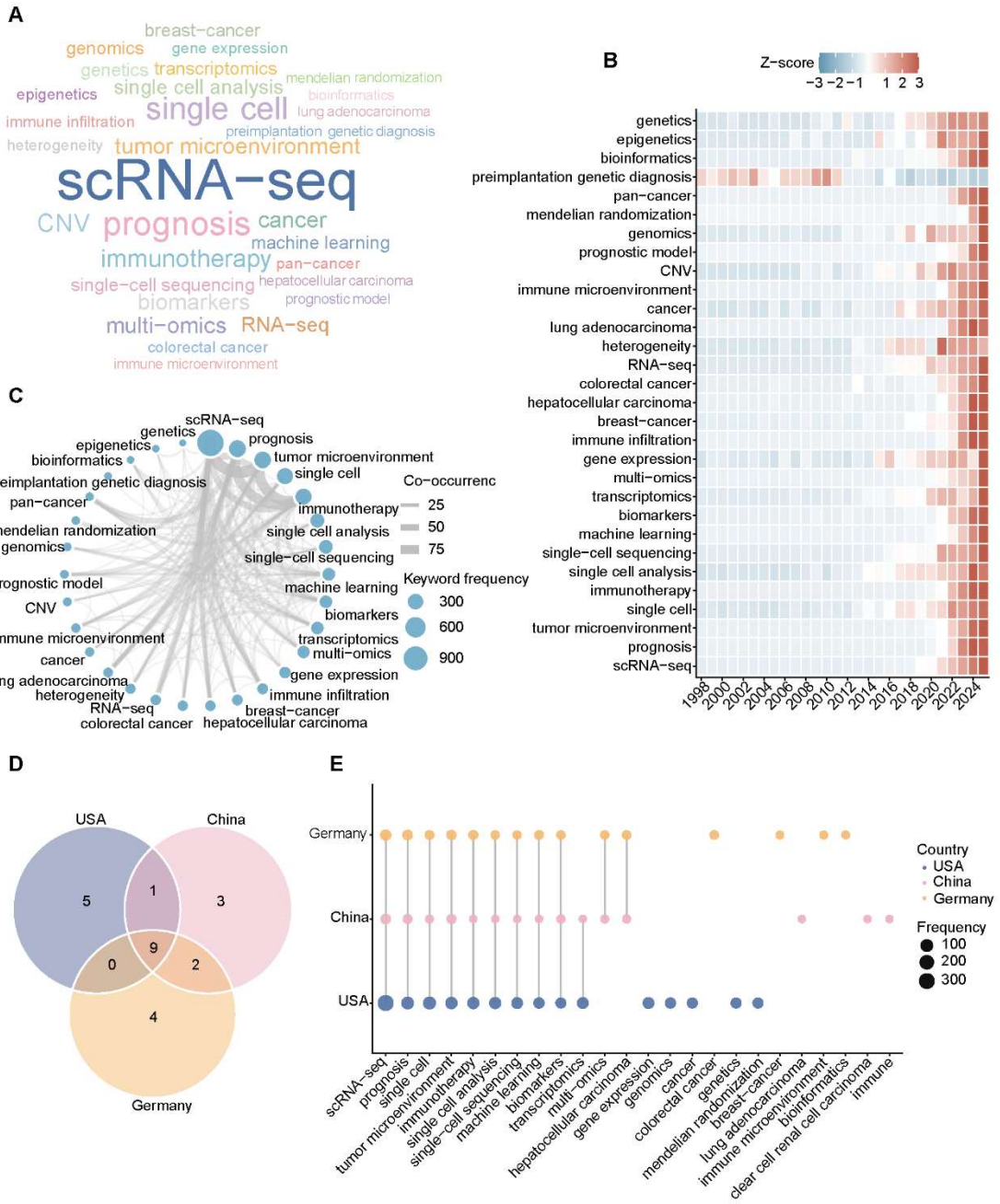


Fig. 3 Keyword Hotspots and Country-Specific Research Themes in Single-Cell Variation Research. **A** Top 30 keywords ranked by occurrence frequency. **B** Temporal trends of high-frequency keywords across publication years. **C** Keyword co-occurrence network showing relationships among major research topics. **D** Overall distribution of high-frequency keywords among the three countries with the highest publication output. **E** Similarities and differences in keyword compositions among the three countries with the highest publication output.

Advances in single-cell sequencing technologies have driven the transition of single-cell genetic variation research toward multi-omics integration

Using the ChatGPT-4o model, we extracted key information from the full text, including sequencing technologies, variant types, disease categories, and species origins. To evaluate the

reliability of the LLM-assisted extraction, 30 publications were randomly selected for manual validation. Averaged across these four information categories, the extraction achieved a mean correctness of 82.50% and a mean completeness of 83.05%. (Table S6; Fig. S5A). These results support the feasibility of ChatGPT-4o for structured information extraction from biomedical full-text publications and are consistent with previous studies demonstrating the potential of LLMs for reliable biomedical information extraction [24, 25].

Next, we extracted sequencing technologies from full-text publications and classified them into five major categories: Single-cell, Bulk DNA, Bulk RNA, Bulk Chromatin, and Spatial transcriptomics (Table S7). The proportion of studies using single-cell sequencing increased markedly from 11.4% in 2013 to 84.07% in 2024, while spatial transcriptomics rose from 0.58% in 2020 to 3.59% in 2024 (Fig. 4A). Among single-cell technologies, scRNA-seq showed the strongest growth, increasing from 37.5% in 2013 to 91.2% in 2024 (Fig. 4B). In contrast, scDNA-seq declined from 62.5% in 2013 and 63.9% in 2014 to 3.6% in 2024, whereas scATAC-seq increased gradually from 0% in 2013 to 5.2% in 2024.

We further analyzed pairwise technology combinations. scRNA-seq + RNA-seq was the most prevalent combination among the top five, increasing from 33.3% in 2014 to 47.3% in 2024. After 2020, scRNA-seq + scATAC-seq increased from 6.3% to 17.2%, while scRNA-seq + spatial transcriptomics rose from 3.8% to 15.6% by 2024 (Fig. 4C). These trends indicate a clear shift from single-omics studies toward increasingly integrated multi-technology approaches at single-cell resolution.

Comparison of the top five sequencing technology combinations across the five most productive countries revealed distinct national patterns. scRNA-seq + RNA-seq was the dominant combination in China, the USA, the UK, and Germany, accounting for 58.1%, 38.1%, 51.3%, and 43.4%, respectively, whereas Japan showed the highest proportion of scRNA-seq + WES at 40%. In the USA, scRNA-seq + scATAC-seq ranked second at 28.1%, while in Germany, scRNA-seq + scDNA-seq ranked second at 30.4% (Fig. 4D). Studies integrating multiple sequencing technologies also tended to receive more citations than single-technology studies, suggesting a potential association between technological integration and greater academic impact (Fig. S5B).

.

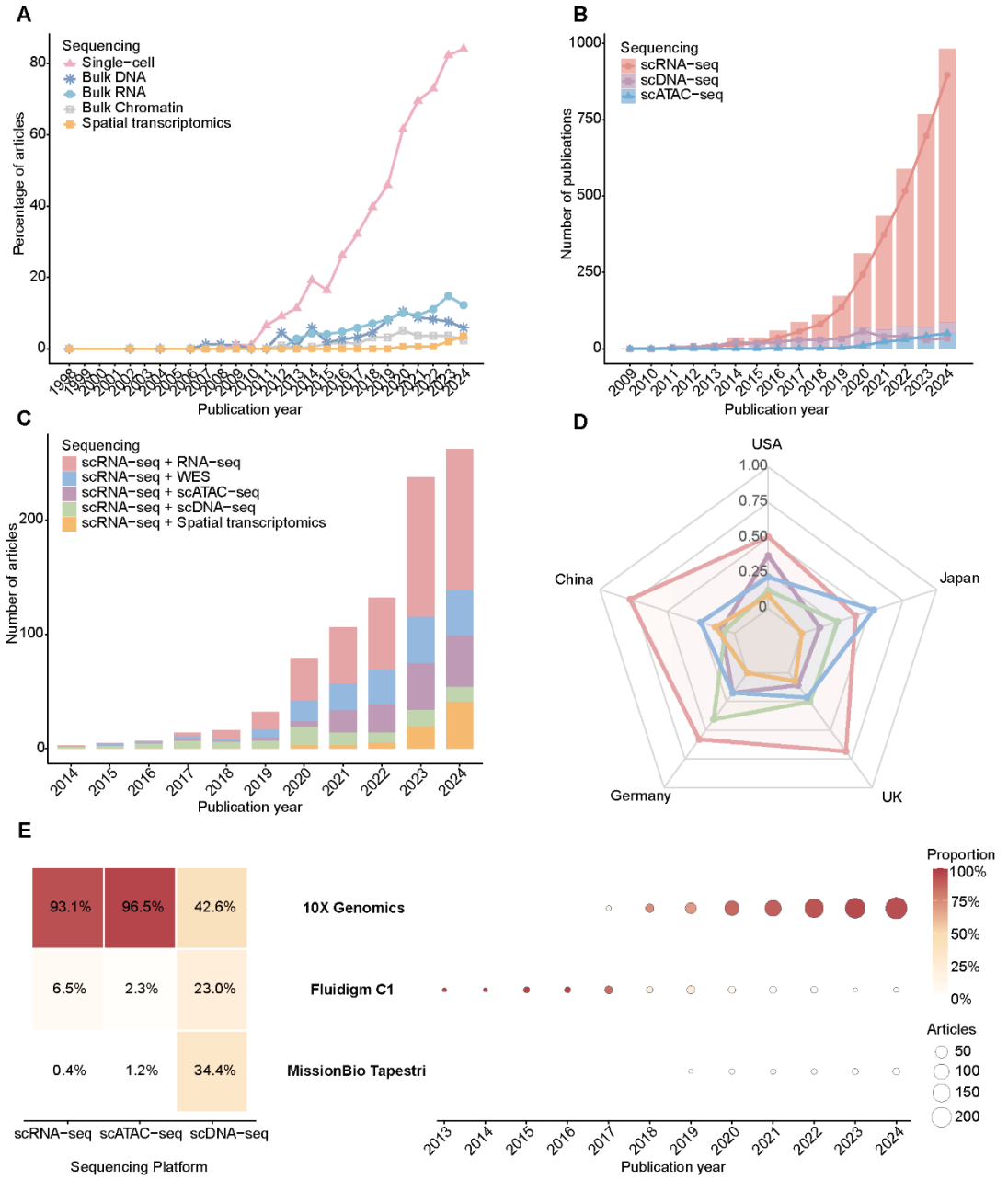


Fig. 4 Sequencing Technologies, Technology Combinations, and Single-Cell Processing Systems. **A** Annual publication trends of different major sequencing technology categories. **B** Annual publication trends of different single-cell sequencing technologies, including scRNA-seq, scATAC-seq, and scDNA-seq. **C** Temporal trends in major pairwise sequencing technology combinations. **D** Differences in sequencing technology combination patterns among the top five countries/regions ranked by publication output. **E** Distribution and temporal trends of single-cell processing systems.

Finally, we examined the single-cell capture and library-preparation systems used across single-cell studies. The three most frequently reported systems were 10X Genomics, Fluidigm C1, and Mission Bio Tapestri. 10X Genomics predominated in scRNA-seq and scATAC-seq studies, accounting for 93.1% and 96.5% of applications, respectively. By contrast, Mission Bio Tapestri showed a higher representation in scDNA-seq studies, accounting for 34.4% of

applications. As a targeted single-cell DNA multi-omics system, Tapestri enables simultaneous profiling of genomic variants and protein expression, making it well suited for analyses of tumor genetic heterogeneity and subclonal architecture[26]. Fluidigm C1 was widely used in earlier single-cell transcriptomic studies, accounting for 92.3% of scRNA-seq applications in 2017, but its utilization declined to 23.8% in 2018. This decrease coincided with the increasing adoption of high-throughput droplet-based systems such as 10X Genomics, which enabled substantially larger numbers of cells to be profiled [27, 28] (Fig. 4E).

Furthermore, Spatial transcriptomics was most frequently combined with scRNA-seq, accounting for 65.8% of spatial transcriptomics studies (Fig. S5C). Among the top three disease categories, cancer was predominant at 84.1%, highlighting the role of spatial transcriptomics in characterizing tumor spatial heterogeneity and the tumor microenvironment (Fig. S5D). Variant analyses were also mainly integrative, with SNV + CNV representing the most common combination at 54.8% (Fig. S5E). Overall, these findings indicate a shift from scRNA-seq-centered single-technology studies toward integrated multi-omics, spatial, and genomic analyses.

Single-cell sequencing technologies are primarily applied in cancer research and focus on tumor microenvironment, prognosis, and immunotherapy

To characterize species sources in studies using scRNA-seq, scATAC-seq, and scDNA-seq, we focused on the three most common categories: Human, Mouse, and Human + Mouse. Human samples accounted for the largest proportion at 61.9%, followed by Human + Mouse at 27.6% and Mouse at 10.6% (Fig. S6A). Over time, mouse-only studies decreased from 19.6% in 2019 to 10.3% in 2024, whereas human-based studies increased from 48.2% to 66.8% (Fig. S6B). These trends indicate an increasing emphasis on human disease samples, while mouse models remain important for mechanistic validation. Consistently, cancer-related single-cell studies were also predominantly based on human samples, accounting for 77.4% among the top three species categories (Fig. S6C).

Based on these findings, we identified the top 20 most frequently studied diseases and grouped them into broader disease categories (Fig. 5A). Single-cell sequencing technologies were predominantly applied in Cancer-related research, accounting for 80.9% of the top 20 diseases (Fig. 5B). Within the Cancer category, lung cancer was the most frequently studied disease, accounting for 19.3% of cancer-related studies (Fig. 5C). The five most frequently studied cancer types all increased rapidly after 2019, with lung cancer showing the strongest growth from 8 publications in 2019 to 81 in 2024 (Fig. S6D), reflecting the expanding application of single-cell sequencing in cancer research.

We next compared the use of scRNA-seq, scATAC-seq, and scDNA-seq across the five cancer types with the highest publication output. scRNA-seq was the dominant technology in all five cancer types, accounting for more than 70% of studies and remaining the primary approach for single-cell variation research in cancer. However, the use of other technologies varied across cancer types. Breast cancer showed a relatively higher proportion of scDNA-seq studies at 23.5% (Fig. 5D), which may be related to its pronounced genomic instability and clonal heterogeneity[29]. scDNA-seq enables direct characterization of genetic alterations, CNVs, and subclonal structures, making it well suited for studying genomic heterogeneity and clonal evolution. In contrast, glioblastoma showed a relatively higher proportion of scATAC-

seq applications at 6.1% (Fig. 5D), potentially reflecting its complex cellular-state heterogeneity and lineage plasticity[30]. scATAC-seq can characterize chromatin accessibility and regulatory changes associated with different cellular states, providing insights into cell-state transitions and tumor heterogeneity[31].

Finally, we extracted major research themes associated with single-cell sequencing applications in disease studies and retained topics occurring more than 10 times. These studies mainly focused on tumor microenvironment, prognosis, and immunotherapy. Prognosis-related studies accounted for the largest proportion at 9.7%, followed by immunotherapy at 8.1%, while heterogeneity-related studies accounted for 2.2% (Fig. 5E; Fig. S6E; Table S8). Temporal analysis showed that immunotherapy, tumor microenvironment, and prognosis all increased markedly after 2021, with prognosis-related studies rising from 4 publications in 2021 to 65 in 2024 (Fig. S6F). Overall, these findings show that single-cell sequencing is increasingly used to characterize tumor composition, microenvironment, cellular heterogeneity, and clinically relevant features, while also supporting translational applications such as prognosis evaluation and immunotherapy research.

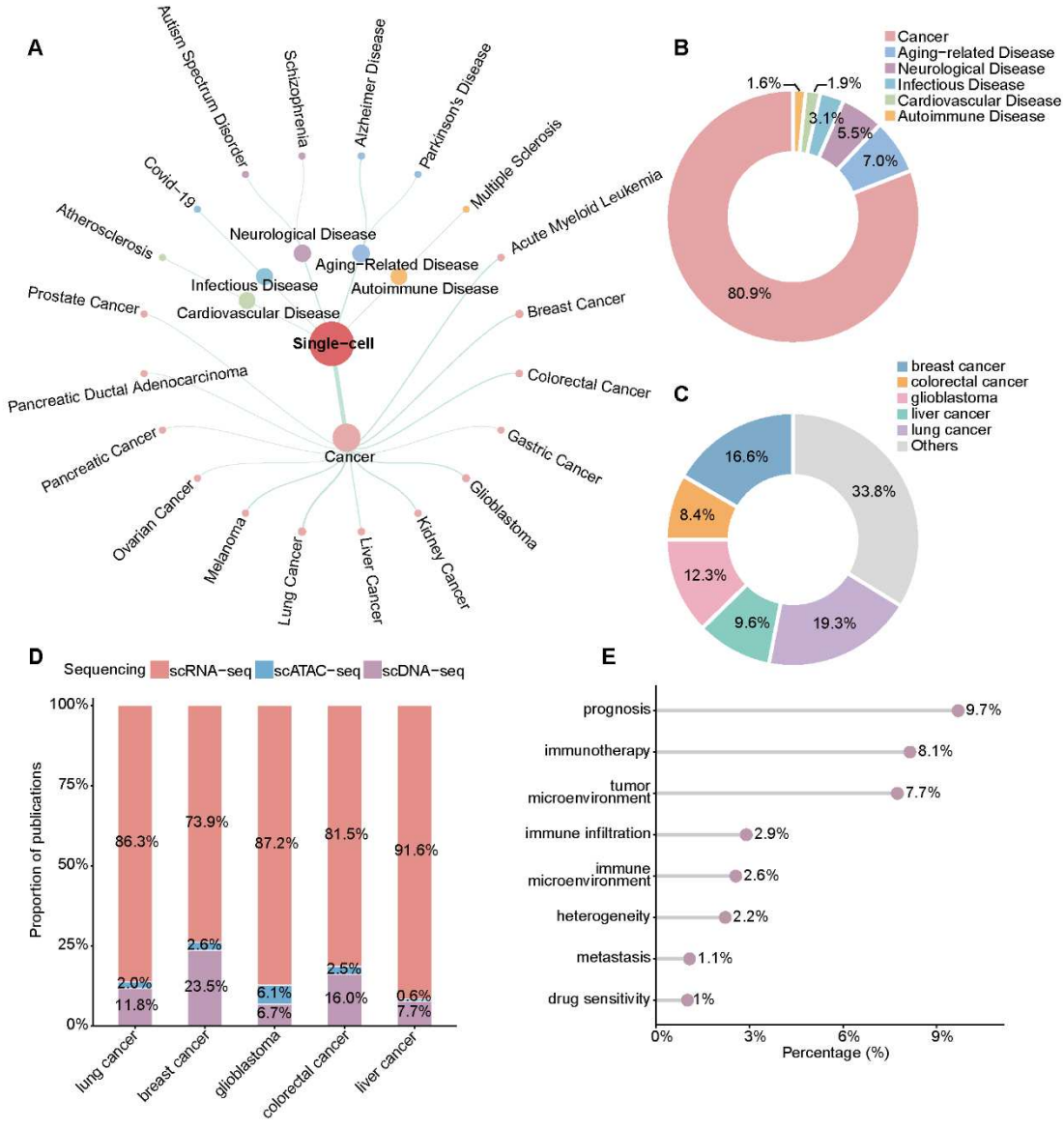


Fig. 5 Disease Applications and Research Themes of Single-Cell Sequencing. **A** Distribution of the top 20 specific disease types ranked by research frequency. **B** Distribution of disease categories among the high-frequency diseases. **C** Distribution of specific cancer types within the Cancer category. **D** Proportions of different single-cell sequencing technologies, including scRNA-seq, scATAC-seq, and scDNA-seq, across the five cancer types with the highest publication output. **E** Distribution of major research themes in single-cell sequencing applications for disease studies (percentages were calculated based on the total frequency of research topics with a frequency greater than 10).

Single-cell genetic variation research is dominated by SNVs and is increasingly shifting toward multi-variant integration and specialized analytical tools

Furthermore, we systematically analyzed the genetic variation types investigated in single-cell genetic variation research. SNV was the most extensively studied variant type, accounting for 51.9% of variant-type-associated studies (Fig. S7A). Its proportion decreased from 54.17%–76.47% before 2009 to 47.96%–60.00% after 2009, although SNVs remained a predominant research focus. In contrast, the proportion of CNV-related studies increased from 8.00%–33.33% before 2009 to 22.88%–42.85% after 2009. This increase is consistent with the strong focus on cancer-related research and highlights the importance of CNV analysis for characterizing tumor heterogeneity and clonal evolution at the single-cell level (Fig. S7B).

Further analysis of co-occurring variant types showed a shift from single-variant analysis toward multi-variant integration. SNV + CNV was the most frequent combination, accounting for 44.0% of variant combination studies (Fig. S7C). Among the five most productive countries, SNV + CNV was the predominant combination in China, the USA, Japan, and Germany, accounting for 59.6%, 38.6%, 46.1%, and 49.0%, respectively. In contrast, the UK showed a higher proportion of SNV + Indels combinations at 36.9% (Fig. S7D). These differences suggest country-specific preferences in variant analysis, potentially related to disease focus, sequencing strategies, and analytical approaches.

We further examined the relationship between single-cell sequencing technologies and variant types. Among scRNA-seq-associated studies, 59.8% involved SNV analysis (Fig. 6A), indicating that scRNA-seq is increasingly used not only for transcriptomic profiling and cell type identification but also for genetic variation analysis. This growing use of single-cell datasets for variant detection has also promoted the development of specialized single-cell variant calling tools.

We further analyzed SNV detection tools used in single-cell sequencing studies. Most studies still relied on conventional bulk-based variant callers, accounting for 97% of reported tool usage (Fig. S7E; Table S9), with GATK being the most frequently used tool at 72.2% (Fig. 6B). Single-cell-specific tools were adopted more recently, with a later median publication year than bulk-based tools (2022 vs. 2021; Fig. 6C), and were mainly reported in JCR Q1 and Q2 journals (Fig. 6D). Although bulk-based tools remain predominant, increasing attention is being given to specialized tools designed to address low sequencing coverage, allelic dropout, and technical noise. These two strategies are currently used in parallel, while future development is likely to focus on further optimization of variant detection methods tailored to single-cell data.

Further analysis across disease categories showed that most disease groups were dominated by SNV-related studies. However, in cancer-related studies, CNV-related studies

slightly exceeded SNV-related studies at 43.5% versus 42.5% (Fig. 6E), highlighting the importance of CNV analysis in tumor subclonal architecture, chromosomal instability, and evolution. Among CNV inference tools, inferCNV was the most frequently used at 58.3%. As inferCNV is designed to infer large-scale copy number alterations from single-cell transcriptomic data [32], its widespread use further underscores the importance of scRNA-seq in tumor CNV analysis (Fig. 6F).

We further compared the disease types most frequently investigated in SNV- and CNV-related studies. Both were predominantly focused on cancer-related diseases, based on their respective Top 10 disease types. Breast and lung cancers were the most frequently studied cancer types in both groups: breast cancer accounted for 18.1% of SNV-related and 19.2% of CNV-related studies, while lung cancer accounted for 17.3% and 16.9%, respectively (Fig. S7F). Overall, single-cell genetic variation research remains centered on SNV analysis, with increasing integration of SNV and CNV profiling and growing emphasis on CNV applications in cancer. Future development will likely focus on specialized variant analysis tools designed to address low sequencing coverage, allelic dropout, and technical noise in single-cell data.

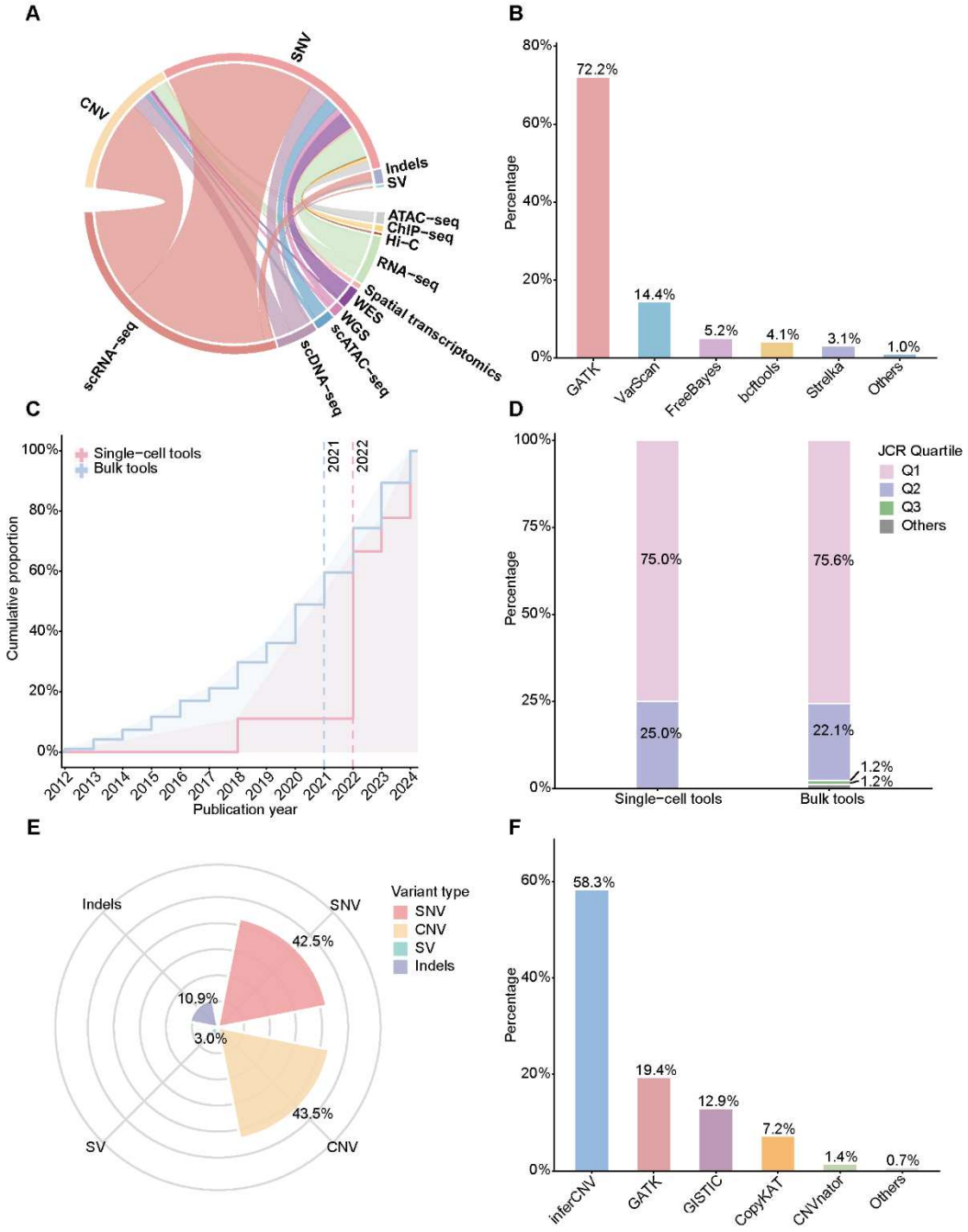


Fig. 6 Associations of Sequencing Technologies with Variant Types, Detection Tools, and Disease Patterns. **A** Association between single-cell sequencing technologies and variant type distributions. **B** Usage proportions of major bulk-based tools for SNV detection. **C** Distribution of publication years for studies using single-cell-specific tools and bulk-based tools. **D** Distribution of SNV-related studies using single-cell-specific tools and bulk-based tools across different JCR quartiles. **E** Distribution of major variant types in Cancer-related studies. **F** Usage proportions of major tools for CNV detection.

Discussion

Using LLM-assisted full-text structured information extraction, this study systematically

characterized the development of single-cell genetic variation research across research trends, international collaboration, technological evolution, disease applications, variant types, and detection tools. Overall, the field has evolved from early technological exploration toward an interdisciplinary landscape integrating multi-omics technologies, disease research, and computational innovation.

First, the rapid growth of publications and international collaboration reflects increasing multidisciplinary and cross-institutional cooperation. Because single-cell genetic variation research involves sequencing technologies, variant detection, computational analysis, and biological interpretation, collaboration across research groups and countries has become increasingly important. The growing contribution of countries such as China further indicates a more diverse and globally distributed research landscape.

Second, sequencing technologies are shifting from single-omics analysis toward multi-omics integration. scRNA-seq remains the predominant technology because of its technological maturity, large-scale datasets, and broad applications in characterizing cellular states and tissue microenvironments [33, 34].However, genetic variation affects genomic, epigenetic, and transcriptional layers that cannot be fully captured by transcriptomic data alone. The increasing use of scATAC-seq, spatial transcriptomics, and multi-technology combinations therefore reflects growing efforts to integrate genomic, epigenetic, transcriptomic, and spatial information[35, 36].

Cancer represents the major disease application of single-cell genetic variation research, reflecting the importance of genetic heterogeneity, clonal evolution, chromosomal abnormalities, and therapeutic resistance in tumors. Compared with bulk sequencing, single-cell approaches can resolve low-frequency subclones and intercellular genetic differences. The prominence of tumor microenvironment, prognosis, and immunotherapy as research themes further indicates that the field is expanding from variant detection toward understanding how genetic alterations influence cellular states, tumor ecosystems, and disease progression[37].

Regarding variant types, SNV remains the predominant focus, whereas the increasing use of combined SNV and CNV analyses indicates a shift toward multi-variant characterization. In cancer-related studies, CNV-focused research is comparable to or slightly more prevalent than SNV-focused research, likely reflecting the widespread chromosomal instability, copy number alterations, and subclonal expansion in tumors[37–39].Previous studies have also demonstrated the value of single-cell CNV analysis for revealing tumor genetic heterogeneity and clonal structure[4, 29].

Variant detection tools also reflect an ongoing methodological transition. Although conventional bulk-based tools such as GATK remain widely used, low sequencing coverage, allelic dropout, amplification bias, and technical noise limit their suitability for single-cell data. Consequently, single-cell-specific tools such as Cellsnp-lite [9], Souporcell[10], SComatic[11], and Monopogen[12], have increasingly emerged, indicating a transition from adapting bulk-based methods toward single-cell-native analytical strategies[39, 40].Several challenges remain. Future studies should improve joint detection of SNVs, CNVs, SVs, and Indels in low-coverage single-cell data, strengthen multimodal integration, and expand applications in tumor clonal evolution, therapeutic resistance, and precision medicine. This will require coordinated

advances in sequencing technologies, computational methods, and clinical sample resources.

This study also has limitations. Emerging studies may appear on preprint platforms before being indexed in Web of Science, potentially limiting coverage of the latest developments. In addition, rapidly evolving technologies and complex multi-omics strategies may introduce omission or classification bias in rare variant types, emerging tools, and research themes. Although prompt optimization and manual inspection were used to reduce extraction errors, LLMs may still have difficulty interpreting complex technical descriptions and distinguishing tool functions, leading to missing information or inconsistent classifications[18, 41].

Despite these limitations, this study demonstrates the potential of LLMs for large-scale structured analysis of biomedical literature. Compared with approaches based mainly on titles, abstracts, and keywords, LLMs can extract multidimensional information from full texts, including sequencing technologies, variant types, analytical tools, and disease categories, enabling more systematic integration of complex biomedical information[17]. With further improvements in extraction accuracy and biomedical knowledge integration, LLMs may become increasingly useful for large-scale literature mining and research trend analysis.

Conclusion

Single-cell genetic variation research is rapidly evolving toward multi-omics integration, multi-variant analysis, and single-cell-specific computational strategies, with cancer remaining its major application area. Our analysis highlights the growing importance of multimodal profiling, integrated SNV–CNV analysis, and specialized variant detection tools in characterizing cellular heterogeneity and clonal evolution. In addition, LLM-assisted full-text information extraction provides an effective approach for systematically capturing and integrating complex methodological and biological information from rapidly expanding biomedical literature.

Abbreviations

API: Application Programming Interface;

CNV: Copy Number Variation;

GPT-4o: Chat Generative Pre-trained Transformer 4o;

Indel: Insertion and Deletion;

JCR: Journal Citation Reports;

LLM: Large Language Model;

NGS: Next-Generation Sequencing;

OLS: Ordinary Least Squares;

PDF: Portable Document Format;

RNA-seq: RNA Sequencing;

scATAC-seq: Single-cell Assay for Transposase-Accessible Chromatin Sequencing;

scDNA-seq: Single-cell DNA Sequencing;

scRNA-seq: Single-cell RNA Sequencing;

SNV: Single-Nucleotide Variant;

SV: Structural Variant;

UT: Unique Web of Science Accession Number;

WES: Whole-Exome Sequencing;

WGS: Whole-Genome Sequencing.

Acknowledgements

Not applicable.

Author contributions

Hechi Zou: Data curation, Formal analysis, Investigation, Writing – original draft, Validation. **Jingjing Zhang**: Formal analysis, Writing – review & editing. **Jiawen Mao**: Validation, Investigation. **Nianling Yang**: Data curation, Formal analysis, Investigation, Validation. **Yongjian Wei**: Data curation, Software, Methodology. **Xiaoran Bai**: Investigation, Data curation. **Hui Zhang**: Investigation, Data curation. **Weiwei Gao**: Supervision, Project administration. **Xianfu Yi**: Supervision, Project administration. All authors read and approved the final manuscript.

Funding

This work was supported by grants from the National Natural Science Foundation of China (No. 32470719 to Xianfu Yi) and the Youth Research Incubation Fund of the School of Basic Medical Sciences, Tianjin Medical University (No. 2023FY01 to Xianfu Yi). The authors gratefully acknowledge Tianjin Medical University for its institutional support.

Data availability

The data analyzed in this study were obtained from the Web of Science Core Collection and publicly available published literature. Processed datasets and materials generated during this study are available from the corresponding author upon reasonable request. The analysis code used in this study is available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Reference

1. Engreitz JM, Lawson HA, Singh H, Starita LM, Hon GC, Carter H, et al. Deciphering the impact of genomic variation on function. Nature. 2024;633:47–57. https://doi.org/10.1038/s41586-024-07510-0.

2. Yu Z, Coorens THH, Uddin MM, Ardlie KG, Lennon N, Natarajan P. Genetic variation across and within individuals. Nat Rev Genet. 2024;25:548–62. https://doi.org/10.1038/s41576-024-00709-x.

3. Bick AG, Metcalf GA, Mayo KR, Lichtenstein L, Rura S, Carroll RJ, et al. Genomic data in the All of Us Research Program. Nature. 2024;627:340–6. https://doi.org/10.1038/s41586-023-06957-x.

4. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011;472:90–4. https://doi.org/10.1038/nature09807.

5. Bowes AL, Tarabichi M, Pillay N, Van Loo P. Leveraging single-cell sequencing to unravel intratumour heterogeneity and tumour evolution in human cancers. J Pathol. 2022;257:466–78. https://doi.org/10.1002/path.5914.

6. Kang J, Lee JH, Cha H, An J, Kwon J, Lee S, et al. Systematic dissection of tumor-normal single-cell ecosystems across a thousand tumors of 30 cancer types. Nat Commun. 2024;15:4067. https://doi.org/10.1038/s41467-024-48310-4.

7. Xue Y, Su Z, Lin X, Ho MK, Yu KHO. Single-cell lineage tracing with endogenous markers. Biophys Rev. 2024;16:125–39. https://doi.org/10.1007/s12551-024-01179-5.

8. Morita K, Wang F, Jahn K, Hu T, Tanaka T, Sasaki Y, et al. Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. Nat Commun. 2020;11:5327. https://doi.org/10.1038/s41467-020-19119-8.

9. Huang X, Huang Y. Cellsnp-lite: an efficient tool for genotyping single cells. Bioinformatics. 2021;37:4569–71. https://doi.org/10.1093/bioinformatics/btab358.

10. Heaton H, Talman AM, Knights A, Imaz M, Gaffney DJ, Durbin R, et al. Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. Nat

Methods. 2020;17:615–20. https://doi.org/10.1038/s41592-020-0820-1.

11. Muyas F, Sauer CM, Valle-Inclán JE, Li R, Rahbari R, Mitchell TJ, et al. De novo detection of somatic mutations in high-throughput single-cell profiling data sets. Nat Biotechnol. 2024;42:758–67. https://doi.org/10.1038/s41587-023-01863-z.

12. Dou J, Tan Y, Kock KH, Wang J, Cheng X, Tan LM, et al. Single-nucleotide variant calling in single-cell sequencing data with Monopogen. Nat Biotechnol. 2024;42:803–12. https://doi.org/10.1038/s41587-023-01873-x.

13. Zhang H, Sashittal P, Karnoub E-R, Jakatdar A, Umeda S, Hong J, et al. Genomic evolution of pancreatic cancer at single-cell resolution. Nat Genet. 2026;58:355–65. https://doi.org/10.1038/s41588-025-02468-9.

14. Williams MJ, Vázquez-García I, Tam G, Wu M, Varice N, Havasov E, et al. Tracking clonal evolution during treatment in ovarian cancer using cell-free DNA. Nature. 2025;647:757–65. https://doi.org/10.1038/s41586-025-09580-0.

15. Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. Nat Rev Genet. 2016;17:175–88. https://doi.org/10.1038/nrg.2015.16.

16. Vandereyken K, Sifrim A, Thienpont B, Voet T. Methods and applications for single-cell and spatial multi-omics. Nat Rev Genet. 2023;24:494–515. https://doi.org/10.1038/s41576-023-00580-2.

17. Dagdelen J, Dunn A, Lee S, Walker N, Rosen AS, Ceder G, et al. Structured information extraction from scientific text with large language models. Nat Commun. 2024;15:1418. https://doi.org/10.1038/s41467-024-45563-x.

18. Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. Research Synthesis Methods. 2024;15:616–26. https://doi.org/10.1002/jrsm.1715.

19. Nykvist B, Macura B, Xylia M, Olsson E. Testing the utility of GPT for title and abstract screening in environmental systematic evidence synthesis. Environ Evid. 2025;14:7. https://doi.org/10.1186/s13750-025-00360-x.

20. Luo X, Rechardt A, Sun G, Nejad KK, Yáñez F, Yilmaz B, et al. Large language models surpass human experts in predicting neuroscience results. Nat Hum Behav. 2025;9:305–15. https://doi.org/10.1038/s41562-024-02046-9.

21. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009;6:377–82. https://doi.org/10.1038/nmeth.1315.

22. Miranda R, Garcia-Carpintero E. Overcitation and overrepresentation of review papers in

the most cited papers. Journal of Informetrics. 2018;12:1015–30. https://doi.org/10.1016/j.joi.2018.08.006.

23. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The Human Cell Atlas. eLife. 2017;6:e27041. https://doi.org/10.7554/eLife.27041.

24. Tian S, Jin Q, Yeganova L, Lai P-T, Zhu Q, Chen X, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform. 2023;25:bbad493. https://doi.org/10.1093/bib/bbad493.

25. Gazzo A, Berardelli S, Biancospino M, Cuollo L, Dei Zotti F, Ferraro E, et al. Benchmarking generative AI tools for literature retrieval and summarization in genomic variant interpretation. Genome Biol. 2026. https://doi.org/10.1186/s13059-026-04237-1.

26. Ruff DW, Dhingra DM, Thompson K, Marin JA, Ooi AT. High-Throughput Multimodal Single-Cell Targeted DNA and Surface Protein AnalysisProtein analysisSingle cellsUsing the Mission Bio Tapestri Platform. In: Ooi AT, editor. Single-Cell Protein Analysis: Methods and Protocols. New York, NY: Springer US; 2022. p. 171–88. https://doi.org/10.1007/978-1-0716-1771-7_12.

27. Xin Y, Kim J, Ni M, Wei Y, Okamoto H, Lee J, et al. Use of the Fluidigm C1 platform for RNA sequencing of single mouse pancreatic islet cells. Proceedings of the National Academy of Sciences. 2016;113:3293–8. https://doi.org/10.1073/pnas.1602306113.

28. Prakadan SM, Shalek AK, Weitz DA. Scaling by shrinking: empowering single-cell “omics” with microfluidic devices. Nat Rev Genet. 2017;18:345–61. https://doi.org/10.1038/nrg.2017.15.

29. Baslan T, Kendall J, Volyanskyy K, McNamara K, Cox H, D’Italia S, et al. Novel insights into breast cancer copy number genetic heterogeneity revealed by single-cell genome sequencing. eLife. 2020;9:e51480. https://doi.org/10.7554/eLife.51480.

30. Dirkse A, Golebiewska A, Buder T, Nazarov PV, Muller A, Poovathingal S, et al. Stem cell-associated heterogeneity in Glioblastoma results from intrinsic tumor plasticity shaped by the microenvironment. Nat Commun. 2019;10:1787. https://doi.org/10.1038/s41467-019-09853-z.

31. Guilhamon P, Chesnelong C, Kushida MM, Nikolic A, Singhal D, MacLeod G, et al. Single-cell chromatin accessibility profiling of glioblastoma identifies an invasive cancer stem cell population associated with lower survival. eLife. 2021;10:e64090. https://doi.org/10.7554/eLife.64090.

32. Schmid KT, Symeonidi A, Hlushchenko D, Richter ML, Tijhuis AE, Foijer F, et al. Benchmarking scRNA-seq copy number variation callers. Nat Commun. 2025;16:8777. https://doi.org/10.1038/s41467-025-62359-9.

33. Stuart T, Satija R. Integrative single-cell analysis. Nat Rev Genet. 2019;20:257–72.

https://doi.org/10.1038/s41576-019-0093-7.

34. Argelaguet R, Cuomo ASE, Stegle O, Marioni JC. Computational principles and challenges in single-cell data integration. Nat Biotechnol. 2021;39:1202–15. https://doi.org/10.1038/s41587-021-00895-7.

35. Lim J, Park C, Kim M, Kim H, Kim J, Lee D-S. Advances in single-cell omics and multiomics for high-resolution molecular profiling. Experimental & Molecular Medicine. 2024;56:515–26. https://doi.org/10.1038/s12276-024-01186-2.

36. Wu X, Yang X, Dai Y, Zhao Z, Zhu J, Guo H, et al. Single-cell sequencing to multi-omics: technologies and applications. Biomark Res. 2024;12:110. https://doi.org/10.1186/s40364-024-00643-4.

37. Gao R, Davis A, McDonald TO, Sei E, Shi X, Wang Y, et al. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. Nat Genet. 2016;48:1119–30. https://doi.org/10.1038/ng.3641.

38. De Falco A, Caruso F, Su X-D, Iavarone A, Ceccarelli M. A variational algorithm to detect the clonal copy number substructure of tumors from scRNA-seq data. Nat Commun. 2023;14:1074. https://doi.org/10.1038/s41467-023-36790-9.

39. Satas G, Zaccaria S, Mon G, Raphael BJ. SCARLET: Single-Cell Tumor Phylogeny Inference with Copy-Number Constrained Mutation Losses. cels. 2020;10:323-332.e8. https://doi.org/10.1016/j.cels.2020.04.001.

40. Sollier E, Kuipers J, Takahashi K, Beerenwinkel N, Jahn K. COMPASS: joint copy number and mutation phylogeny reconstruction from amplicon single-cell sequencing data. Nat Commun. 2023;14:4921. https://doi.org/10.1038/s41467-023-40378-8.

41. Syriani E, David I, Kumar G. Screening articles for systematic reviews with ChatGPT. Journal of Computer Languages. 2024;80:101287. https://doi.org/10.1016/j.cola.2024.101287.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFiles.docx
- SupplementaryTables.xlsx