iSMOD: an integrative browser for image-based single-cell multi-omics data

Weihang Zhang [1,†], Jinli Suo [1,2,3,*,†], Yan Yan [4,5], Runzhao Yang [1], Yiming Lu [1], Yiqi Jin [1], Shuochen Gao [1], Shao Li [1,4], Juntao Gao [4,5,*], Michael Zhang [4,5,*] and Qionghai Dai [1,2,*]

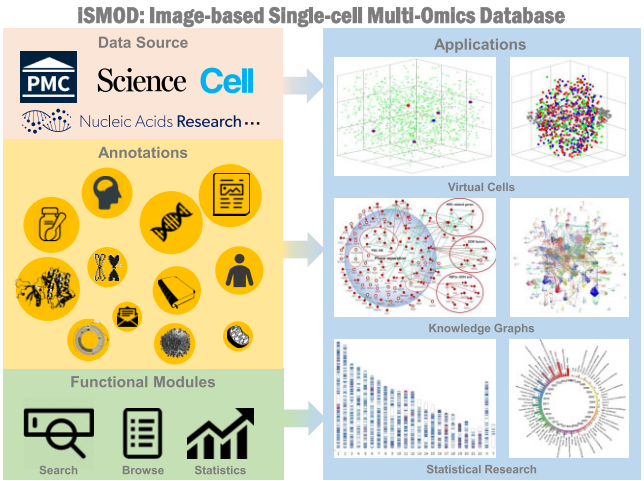
[1]Department of Automation, Tsinghua University, Beijing 100084, China, [2]Institute of Brain and Cognitive Sciences, Tsinghua University, Beijing 100084, China, [3]Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China, [4]MOE Key Laboratory of Bioinformatics; Bioinformatics Division, BNRist; Center for Synthetic & Systems Biology, Tsinghua University, Beijing 100084, China and [5]School of Medicine, Tsinghua University, Beijing 100084, China

Received November 23, 2022; Revised June 09, 2023; Editorial Decision June 20, 2023; Accepted June 26, 2023

ABSTRACT

Genomic and transcriptomic image data, represented by DNA and RNA fluorescence *in situ* hybridization (FISH), respectively, together with proteomic data, particularly that related to nuclear proteins, can help elucidate gene regulation in relation to the spatial positions of chromatins, messenger RNAs, and key proteins. However, methods for image-based multi-omics data collection and analysis are lacking. To this end, we aimed to develop the first integrative browser called iSMOD (image-based Single-cell Multi-omics Database) to collect and browse comprehensive FISH and nucleus proteomics data based on the title, abstract, and related experimental figures, which integrates multi-omics studies focusing on the key players in the cell nucleus from 20 000+ (still growing) published papers. We have also provided several exemplar demonstrations to show iSMOD's wide applications—profiling multi-omics research to reveal the molecular target for diseases; exploring the working mechanism behind biological phenomena using multi-omics interactions, and integrating the 3D multi-omics data in a virtual cell nucleus. iSMOD is a cornerstone for delineating a global view of relevant research to enable the integration of scattered data and thus provides new insights regarding the missing components of molecular pathway mechanisms and facilitates improved and efficient scientific research.

GRAPHICAL ABSTRACT



INTRODUCTION

The progress made in omics has increasingly demonstrated that different types of omics data (e.g. genomics, transcriptomics and proteomics) can crosstalk to ensure mutual regulation and control of various biological processes and diseases. Consequently, research questions can never be fully answered by implementing a single omics type. Integrated multi-omics analysis provides a systematic strategy for understanding the native and altered states of cells in their entirety by threading data from multiple omics fields, greatly expanding our understanding of different cellular processes (1–3).

After decades of research and innovation, the technologies and analysis methods used in the respective omics fields have seen significant advances. The spatial location of genes

*To whom correspondence should be addressed. Tel: +86 010 62788613 839; Email: jlsuo@tsinghua.edu.cn
Correspondence may also be addressed to Juntao Gao. Email: jtgao@tsinghua.edu.cn
Correspondence may also be addressed to Michael Zhang. Email: michaelzhang@tsinghua.edu.cn
Correspondence may also be addressed to Qionghai Dai. Email: qhdai@tsinghua.edu.cn
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

© The Author(s) 2023. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkad580/7223586 by guest on 04 August 2023

and their messenger RNAs is recognized as crucial for their transcription and regulation. In the era of genomics and transcriptomics, fluorescence *in situ* hybridization (FISH) (4) has become an essential method to understand the spatial location of genes and messenger RNAs. Consequently, this method has been widely employed for several decades in the diagnosis of genetic diseases and identification of the genetic aberrations underlying their pathologies (5). Recently, (multi-color) FISH and its high-throughput derivatives, including MERFISH (6), seqFISH/seqFISH+ (7,8), osmFISH (9), split-FISH (10), MINA (11) and others (12,13), have provided a feasible means to effectively verify the chromatin interactions between promoters and enhancers. Most of these methods arise from chromosome conformation capture (3C)-derived methods, such as Hi-C (14) and ChIA-PET (15); such high-throughput acquisition technologies can generate rich genomic and transcriptomic datasets.

By contrast, the spatially and functionally distinct nuclear condensates, such as nuclear bodies, are essential for the regulation and compartmentalization of gene expression in different cell types (16). By self-assembly through phase separation and increasing reaction kinetics, nuclear bodies have been found to contribute to specific nuclear processes and inter-chromosomal interactions (17). Spatial information about the key proteins specific to these nuclear bodies is crucial for understanding their positioning as landmarks to help locate the position of genes and to facilitate gene expression or RNA processing. With the increasing availability of platforms, such as ProteomeXchange (18), Proteomics (19), PRIDE (https://www.ebi.ac.uk/prid), PeptideAtlas (20) and MassIVE (https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp), omics data sharing is rapidly expanding and becoming more popular (21). In fact, datasets on both genomics and proteomics are available (https://ega-archive.org/about/introduction). The availability of such large amounts of data promotes studies on proteomics and the integration of other omics data.

Considering the intrinsic correlation among different omics levels, combining the three-dimensional (3D) location information of DNA, RNA, and key nuclear proteins based on previously described advanced acquisition technologies can help identify the mechanisms of transcription and gene regulation in relation to different diseases. After decades of exploration, there is now a massive amount of rapidly growing yet scattered, valuable multi-omics data on specific topics. It is now possible to integrate these data systematically to accelerate new discoveries. It is, thus, clear that a search engine enabling the aggregation of multi-omics data is required. Surprisingly, all published data on FISH and protein spatial location have not been collected nor organized into databases using comprehensive annotation, classification, description, and statistical analyses, to enable efficient subsequent investigations. Although researchers have recently developed excellent browsers (22–24) for efficient sharing and searching, they do not function with combination studies to integrate the spatial data for DNA/RNA and related proteins.

In this study, we created an integrative browser called iSMOD (Image-based Single-cell Multi-omics Database, https://www.i-smod.com), which collects and browses previously published papers for three types of image-based omics data: genomic (primarily DNA FISH), transcriptomic (mainly RNA FISH), and nuclear proteomic (focusing on key proteins in the nuclear pore complex, subnuclear compartments, nuclear speckles, paraspeckles, nuclear stress bodies, and Cajal bodies) data. iSMOD was built using comprehensive information from the title, abstract, keywords and all experimental FISH or nuclear protein images (currently comprises 23,288 articles and continues to expand) in published research articles available through PubMed. Furthermore, we constructed a search engine that allows users to browse the database with customized search options, such as FISH method, species, gene name or gene location, cancer cell, cell type, allele and protein. The results can be presented in either tabular or graphical forms; the former lists the related papers with publication information (such as title, author, corresponding author, and journal) and experimental sub-figures together with corresponding figure captions and descriptions, while the latter presents the statistics and relationships among the search results in a graphical form.

As an exemplar demonstration of the function of iSMOD, we have described four examples: building knowledge graphs of researchers and various research topics in this field; investigating the molecular mechanism of specific diseases by integrating the FISH of the genes or alleles in a chromatin region; creating a virtual cell to integrate several multi-omics data from different groups; inferring the working mechanism of specific topics from the extracted multi-omics interactions.

As the first integrative browser for image-based multi-omics data focusing on the cell nucleus, iSMOD will be of great benefit to the research community, bringing together the large amount of existing data and research results, helping explore the global scenario, revealing new trends, and inspiring task-specific combinatorial analyses.

MATERIALS AND METHODS

iSMOD data collection and annotation

Article acquisition. For a comprehensive collection of the single-cell multi-omics academic papers, we retrieved all downloadable papers published before April 2023 (regular updates are being made) from Pubmed Central (PMC, https://www.ncbi.nlm.nih.gov/pmc/articles) with the term 'fluorescence in situ hybridization' and a list of nuclear proteins covering the categories such as clastosome, polycomb body, cajal body, perinucleolar compartment, histone locus body, nuclear stress body, lamin, paraspeckle, nuclear speckle, cohesin, nuclear pore complex, nucleolus, condensing, and kleisin. Further, we included additional related papers with the same search keywords from high-impact journals [mainly Science (https://www.science.org/) and Cell (https://www.cell.com/)] that are excluded from PMC. In total, 89 299 papers have been downloaded, among which only those with FISH or nuclear proteins mentioned in the title, abstract, or figure captions of experimental images are reserved.

Keyword library construction. A keyword library was constructed, covering the information at three omics levels to

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkad580/7223586 by guest on 04 August 2023

help organize the database in a systematic way for browsing and searching. To serve users from different research fields, we provided diverse search fields, including species, gene, protein, cell name, cancer, and probe. For each search item, we use crawlers or official FTP to build a relatively complete keyword library.

- The species library is from NCBI (ftp://ftp.ncbi.nih.gov/pub/taxonomy/), with the abbreviation of each scientific name supplemented (e.g., Homo sapiens was abbreviated to H. sapiens).
- The gene library of various species was obtained from http://ftp.ensembl.org/pub/ (25,26). Genes that are named after common daily words (e.g., 'a', 'for', 'fig') or a letter with a number suffix (e.g., 'F3', 'F2') were discarded from the gene list to avoid confusion.
- The cell name library was primarily obtained from ATCC (https://www.atcc.org/en/cell-products/human-cells) and the Human cell markers category in CellMarker (http://bio-bigdata.hrbmu.edu.cn/CellMarker/download.jsp).
- The cancer library was largely obtained from https://www.cancer.gov/types, while the cancer cell list was obtained from CCLE (https://depmap.org/portal/download). For cancers and cells with multiple aliases aside from their scientific name (e.g., APL for acute promyelocytic leukemia), we assigned a unique index.
- The probe library was primarily derived from 'The Molecular Probes Handbook' by ThermoFisher (https://www.thermofisher.cn/cn/zh/home/references/molecular-probes-the-handbook.html).
- The protein library was largely obtained from previous research and reports on nuclear proteins.

We were then able to extract the multi-omics information as textual labels for each article and match the labels with the keyword library.

Automatic figure parsing. Considering that experimental figures often refer to the core points of the article, we extracted the images and their textual information for paper annotation. We then developed an automatic figure parser according to the following steps: (i) Applied optical character recognition (OCR) technology to recognize the text information embedded in a figure, including those in the figure legends (such as A/B/C, etc.). Specifically, we used PaddleOCR (https://github.com/PaddlePaddle/PaddleOCR) implemented by PaddlePaddle based on Tensorflow 1.14, which predicts the location of all text blocks and recognizes the content at high fidelity. (ii) Segmented each figure into panels by adaptive binarization and applying image morphological processing techniques, which is effective in segmenting fluorescent microscopy images. (iii) Designed discriminative image features (described with color moments and color texture moments) and trained a support vector machine (SVM) to filter the non-fluorescent images, such as statistical graphs and ordinary fluorescence images. We excluded manuscripts without figures/panels reserved to prevent them from entering the pool even though relevant textual information, such as FISH, may be contained in the introduction or related context. (iv) Assigned labels for each candidate panel to be the closest OCR-predicted text box with up to two valid characters (either letters or numbers). Moreover, the text inside the panel was regarded as its text description, which generally refers to the related gene, chromosome, or protein. For clarity, we described the workflow of figure parsing in Extended Data Supplementary Figure S5.

Annotation of pairwise interaction information. The extraction of pairwise interaction information was conducted based on the keyword annotations of the massive articles and figures, under semantic guidance of multi-omics interactions. To accurately process millions of related paragraphs from articles in our database, we first identified a set of 'anchors' indicating interactions (including 'interact', 'validate', 'contact', 'loop', 'corroborate', 'promoter' and 'enhancer', along with their derivatives such as the noun and past participle forms) to narrow the search scope. Thereafter, we extracted keywords of genes, alleles, proteins and their aliases from the anchor's neighboring context, and the extracted keywords were judged to have pairwise interactions.

Textual annotation of the database. From an empirical perspective, it was deemed unlikely that the omics items appearing only in the abstract or conclusion section represent the main research object of a paper (particularly if there are no relevant experiments or figures). In addition, if a FISH method or derivative was not mentioned in the abstract, figure caption, or reference context, it may only serve as an introduction or reference and is, therefore, not within our scope of exploration. Accordingly, for an article, we extracted multi-omics terms from the title, abstract, and figure-related text (including the figure caption filtered by the SVM and the context referencing the figures) for annotation, via string matching between these extractive descriptive words and the keyword library. Since string matching with a huge keyword library comprising millions of items is time-consuming, we implemented a rapid matching algorithm adopting the following strategies:

- For each item in the keyword library, two indices were maintained—the string length and number of words.
- Given a descriptive sentence, we parsed it into candidate phrases and searched the library from the items with the matching number of words and letters.
- The search began from the longest items until successful matching.
- For genes, we only searched the gene library of the target species.
- For items that share the same names with common words in English, we counted the number of times that they were mentioned in the manuscript; entities with few appearances were excluded.

iSMOD database and website construction

Article information organization. We organized the dataset via object-oriented programming, with each article, figure, and figure panel serving as an object and annotated with

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkad580/7223586 by guest on 04 August 2023

proper keywords. For an article object, the annotation keywords were obtained from the title and abstract; for a figure object, the keywords were extracted from the caption and reference context matching its label; for a panel object, keywords were primarily obtained from the corresponding text recognized by OCR.

The classes were organized in a hierarchical manner, as illustrated in Extended Data Supplementary Figure S6. First, we defined the `article` class describing the article objects, the data members of which included title, abstract, author(s), correspondence information, author's affiliations, publication time, DOI, journal name and keywords such as species and cancer. Next, within the `article` class, we introduced a member describing its own figures defined by `fig` class, featuring the figure caption, figure ID, image filename, reference context, and various keywords contained in the caption and reference context. Finally, given that an article figure often comprises multiple panels, of which only a proportion are informative for multi-omics studies, we further inherited a 'panel' class from 'fig' to represent the panel objects. Here we added two features, i.e., the label of the panel and the text contents within or about the panel. The keywords, author, and date were also described with corresponding classes, i.e., `kword`, `author` and `data` respectively.

Fortunately, most articles either provide corresponding .xml files in the Pubmed database or html versions on their website, which help retrieve various information through html tags, ID, and references. We extracted the aforementioned content via string parsing, such as searching tags by regular expressions. An example of retrieving the comprehensive information from a paper (27) is shown in Extended Data Supplementary Figure S7.

Database and website construction. Based on the keyword libraries and annotation of the reserved 23 288 articles, we established a two-level database of articles and figures, and sub-databases of authors, publishing dates, corresponding authors, and various keywords, using MySQL 8.0. Furthermore, to ensure facile and extensive use of the database, we constructed a PHP-based iSMOD website on the Elastic Compute Service server provided by Alibaba Cloud, Alibaba Group, China, with a `css` style template, and dynamic behaviors implemented with JavaScript.

We also provide browse and search services for the large database. For customized searches, detailed in the next subsection, one can retrieve a collection of articles via keywords in different fields, and the results will then be displayed comprehensively, including all publishing information of the article and the experimental figure panels. We have achieved rapid queries via the application of a prestored index. In fact, even complex queries involving multiple data fields can be executed within seconds. In addition, we provide the quantitative statistics and graphical visualization of search results in real-time, which is implemented by multi-core CPU using R language and can adaptively select the proper plot type (such as bar plots, pie plots, or circular bar plots) according to the amount of retrieved data. Under the statistics tab of the search result page, one can see the frequency distribution of tens of items (FISH types, proteins, species, cancers, cancer cell names, genes, cell types, cell names, probes, alleles, cell cycles, imaging methods, FISH derivatives, journals, and authors) and frequency evolution across years of species and publishing journals. Simultaneously, iSMOD also collects the interacting pairs in the result list to generate a real-time interaction knowledge graph, which can be further filtered to a sub-graph displaying connections in a specific cell type. For database browsing, we organized the data hierarchically through various types and provide corresponding statistical plots and knowledge graphs for each selected item. Notably, the steps of gathering statistics, drawing statistical plots and knowledge graph based on the search results or item to be browsed are optimized to asynchronous execution and loading, greatly reducing the loading time.

Customized query and result filtering. In addition, iSMOD provides customized query and result filtering, accessed by simply selecting from the combobox or clicking the checkbox in the GUI.

- *Different search options*: By entering the keywords in one or more of the fields in the search boxes of iSMOD, users can retrieve all the relevant published papers from iSMOD. Moreover, iSMOD offers functions for searching the articles related to a given gene or a set of genes within a range of chromosomal coordinates.
- *Comprehensive results*: The search results are displayed in a table with three columns. The left column is a brief list of the publication information, with a hyperlink to the digital object unique identifier (DOI) and corresponding author address. The middle column displays all experimental images obtained by fluorescent microscopy, along with their captions, for which image processing and OCR algorithms are used. The right column is the information set retrieved from the paper, including the species, genes, cancers, cell types, alleles, FISH methods, and other keywords. It is worth mentioning that the corresponding link of the detailed information of genes and proteins, such as position, description, and function, is provided by hovering the mouse over the item, which is achieved by loading a .json file storing links from the Ensembl database (https://ensembl.org, for species-specific genes) and the protein library of National Institutes of Health (https://www.ncbi.nlm.nih.gov/protein, for proteins).
- *Customized filtering*: For the search results, we provide two filtering types for individualized refining. First, the '3D FISH' and 'distance' checkboxes filter the items containing or mentioning 3D FISH or gene distance data, respectively. Second, for queries with chromosomal coordinate ranges, a filter specifying gene names within the range is displayed to reserve a specific subset.
- *Save the search results*: In addition to tabular and graphical presentation, iSMOD prepares a summary, including the number of articles, genes, cell lines, and DNA/RNA FISH types matching the query. Furthermore, we provide an offline text including the list of retrieved information.

Construction of knowledge graphs

All knowledge graphs in our website are generated based on Vis.js (https://github.com/visjs/vis-network) and cover the

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkad580/7223586 by guest on 04 August 2023

relationship network describing author collaboration, author's research field, and entity interaction. Different types of graphs are provided to obtain a complete profile the database and facilitate deep statistical analysis of the relational knowledge of researchers, research topics, and biological interactions from a multi-omics perspective.

Mapping of author collaboration graphs. In the author collaboration graph, the node represents an author and is assigned a weight (i.e. size) proportional to the total citations to all his/her published papers collected in the database. We obtained the citation count for each article in Google Scholar as of April 2023. The edges describe the number of coauthored papers in iSMOD between its associated author pair. Since this database involves nearly 100 000 different authors, the graph only displays author nodes with >1600 citations. Nevertheless, by inputting the names of one or more authors in the multi-line plain-text editing area above the graph, one can obtain a complete collaboration graph associated with the input names, which is expected to promote understanding of the current research status in the field and benefit enhanced cooperation. Besides, the users can customize author collaboration graphs of different FISH methods and their derivatives. Such sub-graphs are built from the subset of articles annotated by the corresponding FISH method.

Mapping of bipartite author-topic graph and its projections. We constructed a bipartite author-topic graph to display the research focus of the authors in iSMOD, with each graph node denoting an author or topic and the edges connecting each author to his/her research topic(s). We defined topic nodes for all FISH methods and nuclear proteins present in iSMOD. The weight of the topic node in this graph was set as the number of articles referring to this method, and that of the author node represents the total citation count in Google Scholar (down-scaled by 10 times to match the range of weight of the method/protein node). To avoid node explosion caused by the huge number of authors, only the author nodes with a weight exceeding 400 are displayed in the bipartite graph. A topic-author edge exists only if the associated author has published at least one article annotated with the corresponding topic. Each edge is assigned a weight proportional to the number of the author's papers in this research area. From all key topics, we can build a global bipartite author-topic graph, as shown in Figure 2A.

To show the author-topic relation knowledge for subfields, we can also generate projection graphs for specific research topics, e.g. the DNA FISH, RNA FISH, nucleic protein, or their sub-classes, according to the annotations of each article. Three representative results are displayed in Figure 2B–D. To maintain consistent configuration among all author-topic graphs/sub-graphs, the nodes in each projection graph contain the same setting as in the global bipartite graph. In the projection graph, all authors (regardless of their citation count) associated with the method/protein are displayed, with the positions of high citation nodes presented in the whole bipartite graph fixed to reveal the unique connection between authors and methods. Correspondingly, the collaborations between newly introduced authors (not included in the bipartite graph) are added as new edges.

Mapping of the entity interaction graph. We construct interaction graphs based on the entity interactions retrieved from articles in iSMOD based on preset anchors. Similarly, the node in the interaction graph denotes an entity that could be a gene, protein, or allele, with its weight proportional to the number of articles referring to this item. An edge represents an interaction between two interacting entities, and the weight is the count of corresponding interactions retrieved from the whole database. For deeper discovery of relational knowledge, we can extend the graph by retaining the genes, alleles, and proteins that are linked with promoters or enhancers.

To serve users from different sub-fields, the graph can be further customized and filtered to display only specific categories by using the function of adding and deleting grouped entities in Vis Dataset, Similar to the author collaboration graph, the filtered sub-graph can also be obtained by inputting the entity name(s) of interest, where only entities with interaction(s) with the input terms, and the enhancer and promoter, are displayed.

Additionally, in the comprehensive representation of customized search results or papers related to the selected item, the interactions in the list are gathered with their corresponding cell type annotations to form a .json file, which is used for the construction of the real-time entity interaction graph. Employing the cell type(s) as a checking label, the sub-graph on a specific cell type can be displayed by selecting the item in a drop-down menu, promoting cell-specific interaction research on multi-omic entities.

The module identification and role classification of the graphs. Guimerà and Amaral have shown that a complex graph can be divided into functional modules; the nodes with varying features and importance can then be categorized into different roles according to their connection pattern intra- and inter-modules (28). By extracting the module and role information in a complex network, either author collaboration or entity interaction graph, one can obtain a coarse-grained version that can better help researchers isolate important clusters or communities and offer intuitive insights into academic leadership and research foci. We utilized the `rnetcarto` package (https://github.com/cran/rnetcarto) in R language to respectively perform module identification and role classification for the graphs. In the graphs describing author cooperation and entity interactions, we used distinct node colors for different modules, and different shapes to distinguish their roles in the graph. For the bipartite author-method graph, we set the option 'bipartite' in the `rnetcarto` to true to perform analysis on authors only.

Metrics of the graph. In the study of the social network of scientific collaborations, quantitative measures are provided to depict the degree distribution and connectivity of complex graphs (29). As for our proposed graphs, we calculated the average and maximum degree of the author collaboration graph; the following formula derives the clustering

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkad580/7223586 by guest on 04 August 2023

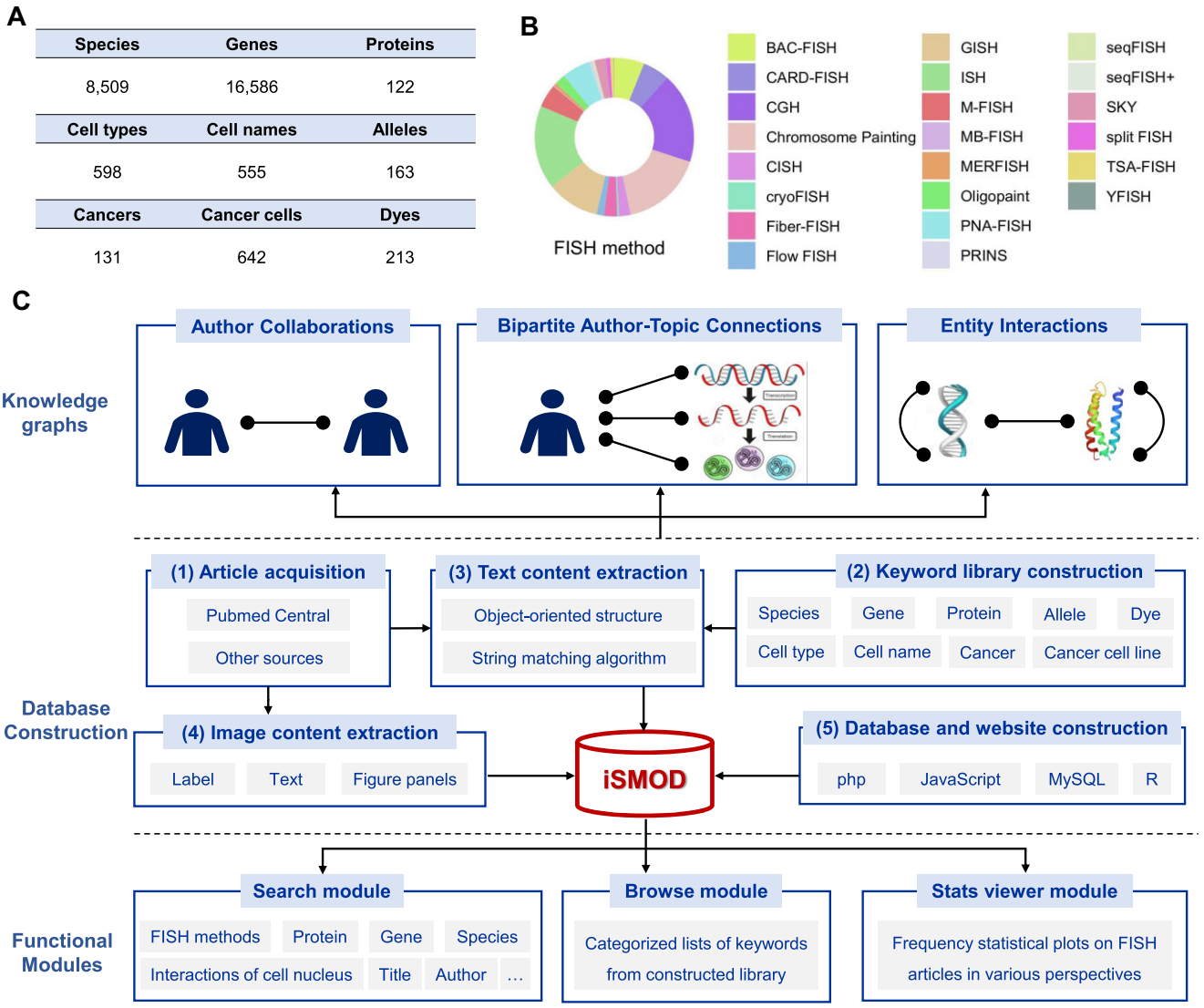


Figure 1. Construction and functions of iSMOD. (**A**) lists the numerous attributes (such as species and gene data) in iSMOD and the corresponding number of entries. Using published literature on various FISH derivatives (**B**) and nuclear proteins, iSMOD extracts relevant figures and key information from multiple perspectives and provides three main functional modules: search, browse, and statistics, as well as three types of knowledge graphs highlighting the connections between authors and entities. The steps are sequentially numbered in (**C**).

coefficient of a node v_i in the complex graph (30):

$$C(v_i) = \frac{2|\mathrm{U}_{v_j, v_k \in N(v_i)} \{\langle v_j, v_k \rangle\}|}{|N(v_i)|(|N(v_i)| - 1)} \quad (1)$$

where $N(v_i)$ is the set of all nodes with connection to v_i. $\langle v_j, v_k \rangle$ is 1 if the nodes v_j and v_k are connected, otherwise it is 0. $|N(v_i)|$ is the size of the set $N(v_i)$. The clustering coefficient represents the actual number of edges formed by the set of immediate neighbors of the node divided by the number of possible edges formed by the set. The average clustering coefficient characterizes the property of dense connection and can be derived by averaging the clustering coefficient of all nodes in the graph. The density of a knowledge graph proposed by Otte and Rousseau quantitatively depicts the aggregation from another perspective as (31)

$$\Delta = \frac{2G}{N \times (N-1)}, \quad (2)$$

where N is the number of nodes and G is the number of edges. Finally, the Floyd's algorithm (32) is performed to calculate the diameter of the proposed networks, i.e., the maximum value of the shortest distances between any two nodes in the graph that are connected directly or indirectly.

Construction of 3D virtual cells

We demonstrate iSMOD's 3D reconstruction annotation and data integration capabilities through the four virtual cells shown in Figure 4B and E–G, which have been visualized in an interactive manner on the homepage GUI of iSMOD. Specifically, the construction of virtual cells

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkad580/7223586 by guest on 04 August 2023

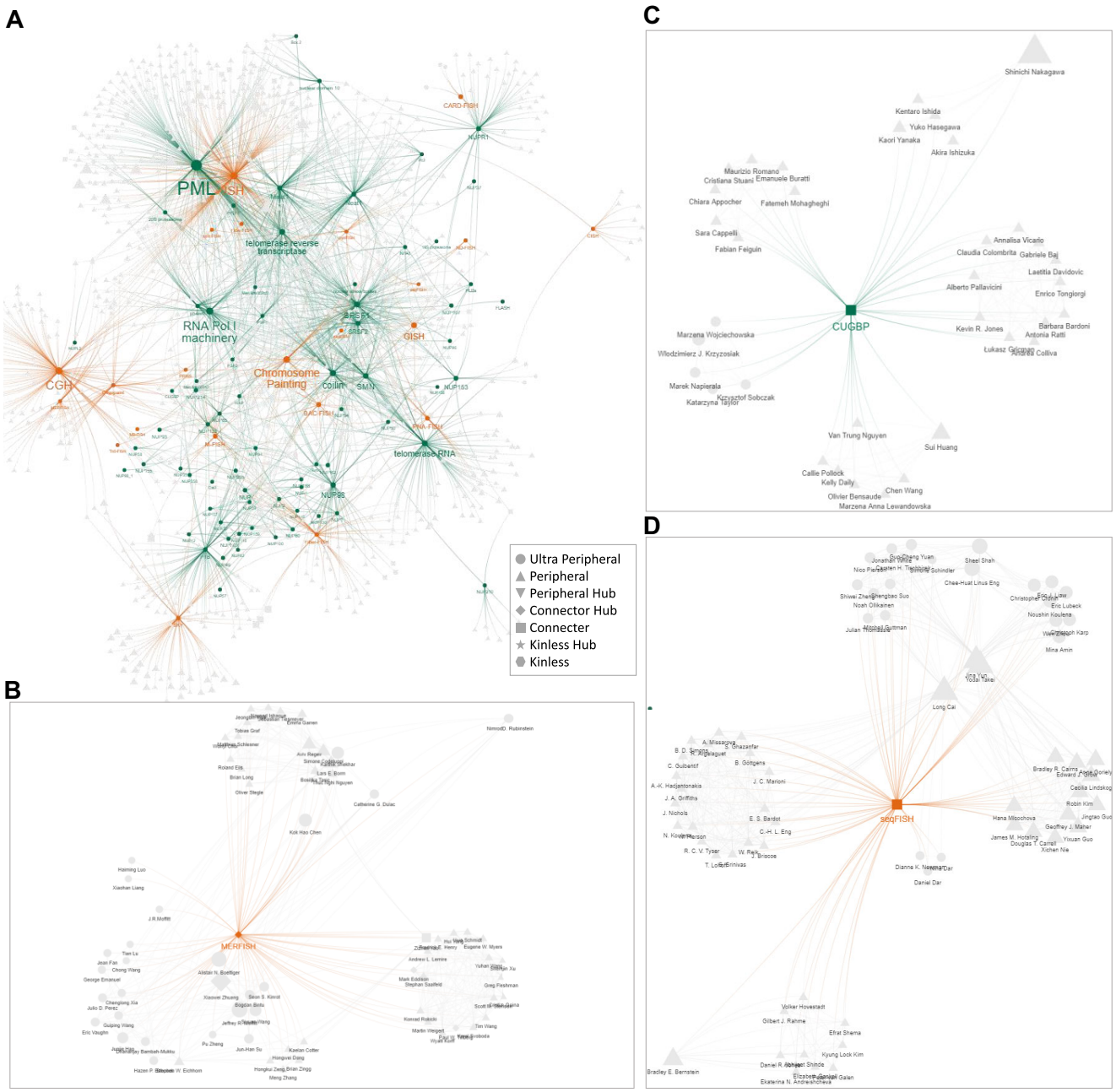


Figure 2. Bipartite author–topic graph and its projection graphs. (**A**) The bipartite author–topic graph built from the iSMOD database. (**B**–**D**) The projection author–topic graphs for DNA FISH (MERFISH), a nuclear protein (CUGBP) and RNA FISH (seqFISH), respectively. The green nodes represent the nuclear proteins, the orange nodes represent the FISH derivatives, and the gray nodes represent the authors. The roles of the nodes are depicted by their shapes.

can be described in two steps: 3D data collection and 3D modeling.

3D data collection. All source data in the exemplar models rely on the compound search and comprehensive annotation of iSMOD. For the distribution of the same gene at different stages, we input the gene name or chromosome interval of interest (e.g. `Xist`) in the 'Gene' text box of iSMOD, select 'FISH type' in the 'Other Keywords' options, and select the 3D option in the filter box of the search result page. As the figures and captions provided by iSMOD can indicate whether the manuscript contains images of related genes, potential candidates that may provide 3D FISH experimental images or coordinate lists can be obtained. Due to limited data availability, 3D images or coordinate data typically appear in supplementary materials/source data attached to the article or are provided directly by the author. For example, in the two pieces of source data integrated in Figure 4B, the article by Takei et al.introduces its data repository (https://zenodo.org/record/3735329) in the

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkad580/7223586 by guest on 04 August 2023

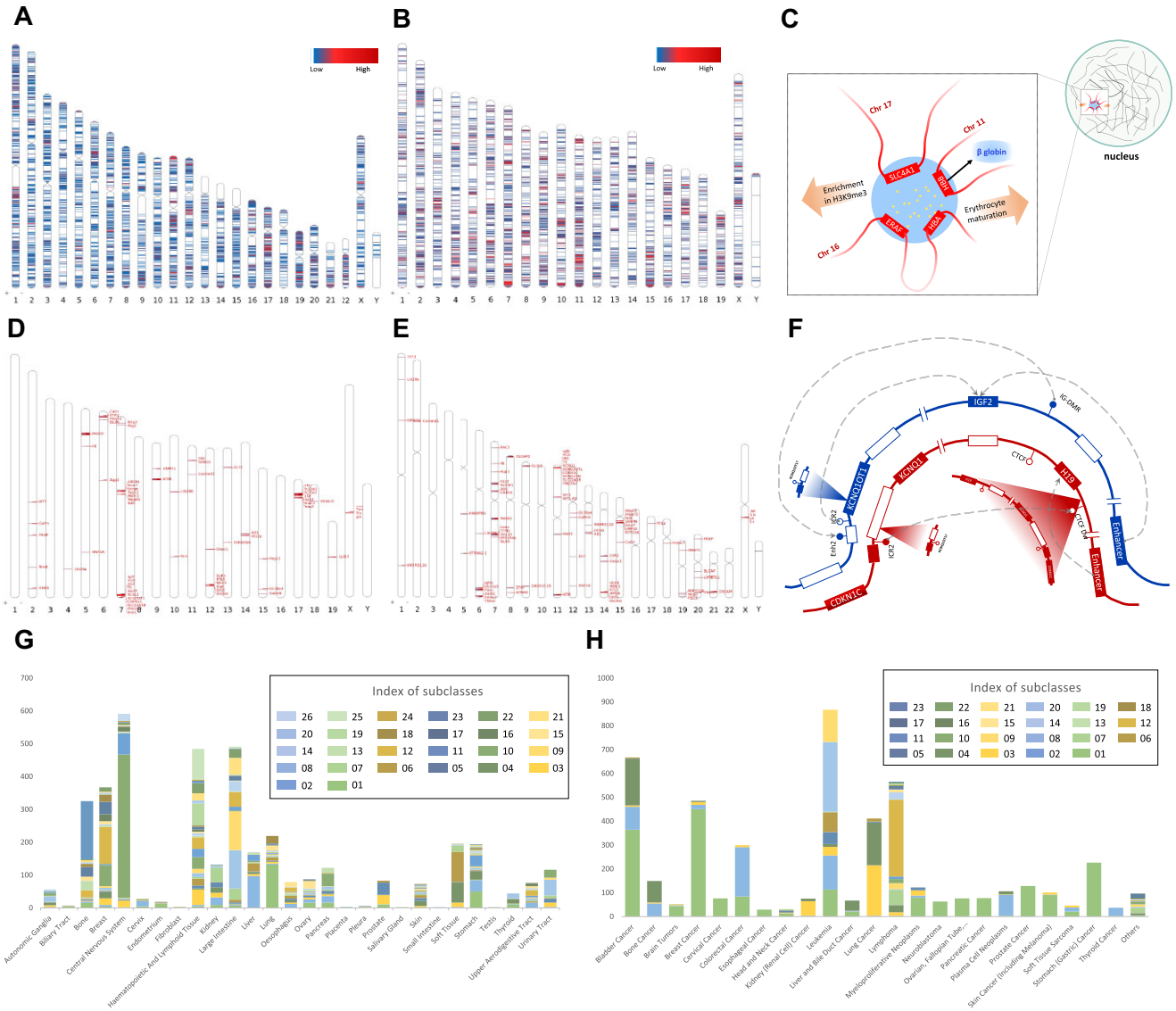


Figure 3. Examples of molecular mechanism investigations for diseases, based on the integrations of FISH data for the genes or alleles in a given chromatin region. (A, B) Distribution of genes in human (**A**) and mouse (**B**) chromosomes studied in iSMODdatabase. (**C**) The mechanism inferred from the retrieved articles studying gene `HBB` in region `Chr11:5225464-5269945`. (D, E) Distribution of alleles imaged by FISH in human (**D**) and mouse (**E**) chromosomes. (**F**) The graph produced by integrating results from iSMOD on the Beckwith–Wiedemann and Silver–Russell syndromes, which are closely related to alleles in the `11p15.5` region. The red boxes represent the maternally expressed genes, whereas the blue boxes represent the paternal genes. (**G**, **H**) Stacked bar plot for cancer cell lines (panel G, categorized by organs) and for cancers (panel H, categorized by general cancer types).

'Data Availability' section (33), while the article by Shiura et al.provides the original 3D image files in the email communication with us (34). The former contains the voxel-wise measurement of 3D coordinates of multiple genes and provides sufficient information to allow the labelling of specific genes, such as `Xist`, and integration with data from other articles. Specifically, we derived the convex hull of the coordinate set of all genes in each cell and regarded its centroid as the center of the cell. We then obtained the location of the `Xist` with respect to the cell center based on the given voxel size. For the files provided by the latter, the vesicle labeling and cell segmenting functions in Imaris were applied to generate a three-dimensional coordinate list of `Xist` in cells at different developmental stages. For the joint visualization of multiple multi-omic entities in the same type of cells, we employed a similar method to obtain accessible data from iSMOD, with 'MERFISH' selected in the 'Method' options and the 3D filter enabled. As a result, several articles guided us to obtain MERFISH imaging data of the mouse brain receptor map from the official website of Vizgen, Inc. (https://info.vizgen.com/mouse-brain-data), which aided us in building the model shown in Figure 4 E. The model shown in Figure 4F and G is derived from the article by Su et al. (35) in search results that provides source data in the "Data and Code Availability" section (https://zenodo.org/record/3928890), which contains a 3D coordinate list of genes, chromosome coordinates, transcriptional information, and distances to nucleoli or

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkad580/7223586 by guest on 04 August 2023

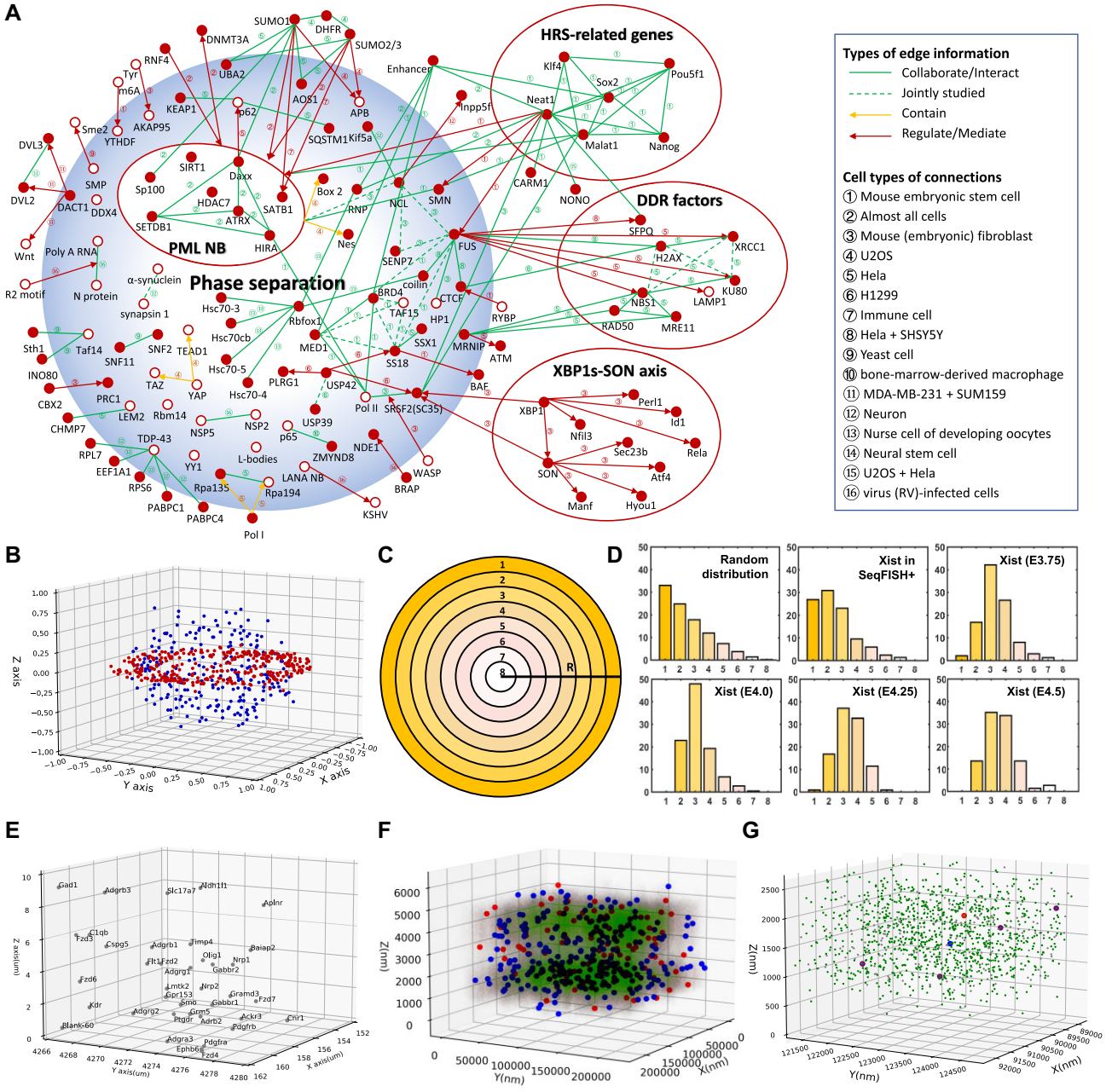


Figure 4. Examples of exploring the working mechanism under biological phenomena and integrating 3D omics data from different sources. (**A**) Knowledge graph of entity– connections in terms of phase separation. Different between-node relationships are distinguished with edges of different line types, and the numbers indicated on the edge represent the index of the corresponding cell type or cancer cell line, listed in the legend. (B–G) Integration and analysis of single-cell multi-omics data, focusing on genomic data with the spatial distributions of `Xist` in different retrieved references as one example (panels B–D) and spatial transcriptomic data with MERFISH in mouse brain cells (panels E–G). (**B**) Based on the 3D coordinates of many genomic loci from seqFISH, the location of `Xist` (shown in blue) in other phases is added after the normalization of the 3D coordinate system. (**C**) Schematic showing the analysis of the 3D position of `Xist`, using shells of equal radius division. (**D**) The assumed random distribution of genes in different shells, analysis of the 3D positions of `Xist` in various shells in seqFISH (33) and analysis of the 3D positions of `Xist` in various shells during different developmental stages (E3.75, E4.0, E4.25 and E4.5) in Shiura *et al.*'s study (34). (**E**) The 3D visualization of MERFISH data in mouse brain cells. (**F**) Multi-omics visualization of the iSMOD integration of single IMR-90 cells in 3D space. The genes that are not transcribed are shown in green, whereas those transcribed are depicted in gray. Red and blue dots denote nucleoli and speckles, respectively. (**G**) Zoomed-in view of (F) with coordinates region *x*:88737-92337, *y*:121237-124837, *z*:0-2679 (unit:*nm*). The genes that overlapped with `homeobox` (`HOX`) gene family on chromosomes are marked in purple.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkad580/7223586 by guest on 04 August 2023

speckles. Notably, we reserved the coordinates with distance to nucleoli or distance to speckles <5 nm, which can be considered as the positions of nucleoli and nuclear speckles. In addition, in the zoomed-in model shown in Figure 4G, We highlighted the genes that overlap with the HOX genes family (HOXD1, HOXD3, HOXD4, HOXD8, HOXD9, HOXD10, HOXD11, HOXD12, HOXD13, HOXC4, HOXC5, HOXC6, HOXC8, HOXC9, HOXC10, HOXC11, HOXC12, HOXC13, HOXA1, HOXA2, HOXA3, HOXA4, HOXA5, HOXA6, HOXA7, HOXA9, HOXA10, HOXA11, HOXA13, HOXB1, HOXB2, HOXB3, HOXB4, HOXB5, HOXB6, HOXB7, HOXB8, HOXB9, HOXB13) on chromosome coordinates as purple dots.

3D modelling. We have built the interactive 3D virtual cell models in the "3D Virtual Cell" column of the iSMODhomepage. Specifically, we performed normalization on the coordinates of the points in each virtual cell and then recorded them in several arrays as a .json file according to different categories (e.g., gene, transcription, nucleoli, and speckles), with the category name serving as keys. In particular, in the model shown in Figure 4 E, each point object contains a key-value pair for gene name, while in the model shown in Figure 4F and G, the innumerable gene and transcription point objects contain a key-value pair for transparency, which are usually relatively small. Based on the API of the Lufylegend.js engine (http://lufylegend.com/api/en_US/out/index.html) for model plotting and Ajax technology for loading .json files, the drawing of coordinate planes and points is carried out on the Canvas element. For convenient use of the models, we set up a mouse event enabling users to change the perspective by dragging the mouse. Further, by clicking the corresponding title, a new page with the enlarged version of the virtual cell is accessed, where one can further filter different omics tags by making part of the points invisible on the canvas. In addition, a range slider is provided to zoom in/out on the model, which is achieved by redrawing each element on the canvas with a updated size.

RESULTS

Data sources and functions of the iSMOD database

iSMOD has collected 23 288 papers on FISH and nuclear proteins in the life sciences and medical fields from PubMed. For each paper, we conducted systematic processing for comprehensive textual annotation, including FISH types (DNA or RNA), proteins, species, gene names, cancers, cell types, cell names, cell cycles, dyes, FISH derivatives, and other keywords or search terms. Additionally, the genes and proteins were recorded with the links to their detailed information in the Ensembl gene database (https://ensembl.org) and the protein library of National Institutes of Health (https://www.ncbi.nlm.nih.gov/protein), respectively. The distributions in terms of different labels and FISH techniques are shown in Figure 1A and B, respectively. Specifically, we extracted the textual annotations (i.e., keywords) of each article from its publishing information, title, abstract, and figures with their embedded texts, captions, and related descriptions parsed automatically, as demonstrated in the middle block of Figure 1C.

To enable researchers in related fields to easily utilize the tool and thus expand the flexible and diverse use of iSMOD, we developed a user-friendly graphical user interface (GUI) based on php, as shown in the lower block of Figure 1C. The GUI has three key modules. (i) The 'Search' module supports customized searching of the repository with different annotated fields such as the method, species, gene, and users can use either a simple query search with one field or a more complex search using multiple fields. (ii) The 'Browse' module allows users to browse through the database based on different categories, including species, cancer, gene, protein, probe, allele, journal, and author, and provides statistical analysis of the selected item (e.g., Myc in gene module). (iii) The 'Stats viewer' module provides a comprehensive graphical illustration of the categorical distribution of the search results or selected item in the browser, by plotting the qualitative statistical figures of the database from different perspectives. The GUI is accessible in Windows, Linux, MacOS and Android systems and supports multiple browsers and screen resolutions. The user interface for the three key modules has been demonstrated in Extended Data Supplementary Figure S1, Extended Data Supplementary Figure S2, and Extended Data Supplementary Figure S3, respectively. For other functions, please refer to the user manual (https://www.i-smod.com/user_manual.pdf), which provides detailed documentation and step-by-step instructions for the usage of iSMOD.

Such a large database, together with comprehensive annotations of the author, research topics, and entity interactions, is bound to form a huge, complex associated network. iSMOD provides tools for constructing knowledge graphs to analyze the key insights buried in such a huge network. Generally, we built three types of graphs for integrative study: author–author connection (i.e. collaboration), author–topic connection, and entity–entity connection (i.e. biological interaction), as shown in the upper block of Figure 1C.

Statistical characteristics of iSMOD

With such a large repository and comprehensive annotations, iSMOD provides multi-dimensional statistical plots, revealing the focus and trends of multi-omics research. The bar plots in Extended Data Supplementary Figure S4A illustrate the proportions of different species, genes, and cancers that can be viewed via the 'Stats viewer' GUI. In addition, the distribution is displayed in Extended Data Supplementary Table S1 with only the top 10 entries reserved. The sorting and distribution can help identify the hot topics in this field. For example, from the frequency distribution of the articles based on specific species, one can see that research on humans and mice has been the predominant focus, and the use of Drosophila and budding yeast has also been relatively common in FISH analysis. Coupled frequency plots are also provided in Extended Data Supplementary Figure S4B. By enumerating the research papers on different species over the years, the development trends for species studied using multi-omics can be assessed. Additionally, the proportions of different imaging methods and different omics levels are plotted in Extended Data Supplementary Figure S4C. For instance, one can roughly

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkad580/7223586 by guest on 04 August 2023

determine how many genes have allele-specific FISH data using the frequency plot for allele names. Overall, we can obtain tens of informative statistics and leave more open for users.

GUI of the search engine

Customized searching. To make iSMOD more user-friendly, we have constructed a website that has a user-friendly GUI for customized searching, as shown in Extended Data Supplementary Figure S1. The interface provides various aforementioned search fields and supports switching among three display modes for the comprehensive search results (tabular listing, graphical plots, and knowledge graph) implemented in the GUI of stats viewer. Moreover, options are available for filtering 3D data and saving the annotated search results and publication information in a summary document; see methods in Section 'Customized query and result filtering' for details.

Repository browser. While the 'Search' module offers a rapid and facile means for users to view multi-omics articles with specific titles or attributes, the 'Browse' module comprehensively organizes all data from iSMOD into different perspectives to meet the demands for further exploration in diverse backgrounds. A browser with a GUI was constructed to provide rapid access to the massive amount of categorized data and organizations, as shown in Extended Data Supplementary Figure S2. As a user switches among the categories, all entries of this category in iSMOD are displayed as a clickable list, and a circular bar plot of the top 100 frequency entries in the category is displayed, which intuitively characterizes the focus of related studies to some extent. By selecting an entry of interest in the list, a new page similar to the search results is displayed to provide an overview for all article information and statistics in the sub-database containing this entry. The offline summary document is also available for download from the summary module in this page.

Stats viewer. The website has a stats viewer to display the search results in three modes: The tabular listing shows the information of each item matching the query, whereas the graphical plot and knowledge graph (interaction network) provide statistical depictions rendered from the search results on another page, resulting in a more intuitive overview of the query. Users can switch among the three modes using the radio button on the results page. On browsing the repository, the users can also see the statistics of the papers related to the selected item using similar graphs.

The graphical plots are generated online using R, and the knowledge graphs are displayed using Vis.js (https://github.com/visjs/vis-network), where colors indicate the modules (similar to a community), and shapes indicate the node functions.

Exemplar Applications

Research profiling of the image-based single-cell multi-omics field. One of the core functions of iSMOD is to generate a global overview of the topics of the user's interest, to assist with tasks such as quickly conducting a scoping review, identifying hot topics or the right collaborators, or locating a specific reference. A huge complex network exists among the comprehensive annotations of tens of thousands of papers in iSMOD, and a proper and flexible graph can reveal important insights.

Scientific cooperation is of great significance for knowledge sharing, resource/information sharing, and advancing innovations (36). This notion can be traced back to Price, the inventor of scientometrics, who proposed the concept of an 'invisible college' and provided effective methods and models for studying the relationship of collaboration (37,38). The rise of social network analysis methods has further promoted the development of research on author collaboration (39,40). To this end, iSMOD provides a knowledge graph of author collaboration (Extended Data Supplementary Figure S8) with advanced filters and subgraphs corresponding to various FISH methods, which are expected to improve collaboration in the following aspects. First, iSMOD depict the scientific collaboration map in different topics of multi-omics research by providing measurements of the clustering coefficient, network diameter (29), or density (31) (see Methods in section 'Metrics of the graph'), which are beneficial for researchers to break down the barrier of collaboration and review the current research status. In Extended Data Supplementary Table S3, we provide the above indicators of the author collaboration graph and the corresponding sub-graphs of the FISH methods currently provided by iSMOD, reflecting the distinguished research practices in different fields. Second, iSMOD produces suggestions regarding author collaboration groups by labeling each node in the collaboration knowledge graph with an advanced method (28). Specifically, the nodes assigned the same labels are associated with the same affiliation or close collaborations. Third, iSMOD measures the experienced experts (large-size nodes), key authors (large-degree nodes), and the frequency of collaboration (mean degree of the graph) (29) in the scientific relationship. Exemplar data on the provided graphs are also displayed in Extended Data Supplementary Table S3. Further, iSMOD analyzes the role of researchers in collaboration, where the centrality measurement methods have the potential to characterize the role of each author in constructing the collaboration graph (41). However, we adopted a more refined classification approach proposed by Guimerà and Amaral (28) for the various graphs provided by iSMOD, where authors (or other nodes) are classified into seven types of roles, indicating the frequency and preference of his/her collaboration with the inter-group and intra-group counterparts (see methods in Section 'The module identification and role classification of the graphs'). In this way, it is expected to effectively improve the efficiency of novice researchers in grasping the development of the field and conducting literature retrieval. Finally, the sub-graph generated through customized queries meets specific needs, such as identifying potential mentors/collaborators or conducting literature research. Specifically, by customizing the graph with arbitrary input, one can further obtain a graph of a narrower scope to get a quick overview of the research status for a group of interest and the enclosed relational knowledge.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkad580/7223586 by guest on 04 August 2023

We provide similar advanced filtering capability for the discovery of new mechanisms by combining relevant knowledge in the multi-omic entity interaction graph, which is generated based on the interaction information of all articles included in iSMOD. It is worth mentioning that we have also embedded a real-time interaction graph in the search result page to describe the connection of multi-omic topics that may be included in the current query, and a filter of cell types is then generated for the exploration of cell-specific interactions. An exemplar application is demonstrated in the subsequent sub-section.

For a global picture of iSMOD in terms of researchers and research topics, we can generate a bipartite graph between authors and topics (e.g. FISH methods or nuclear proteins) (Figure 2A), which serves as a means to comprehensively characterize the state of research in the multi-omics fields. Two different colors are used to encode and distinguish categories. The established methods and research hotspots are highlighted with large font sizes, e.g. PML, Chromosome Painting, and CGH. Considering that some of the new techniques gain momentum given their unique features, the dynamics of the hot topics may change, and regular updates for iSMOD will help identify the recent research trend.

To provide the statistics of a specific research focus at a finer scale, a projection of the bipartite graph can be generated by selecting a specific topic. For example, here, we selected an exemplar instance from each image-based single-cell multi-omics and showed the respective projections. Figure 2B and D present two important high-throughput methods for genomics—MERFISH and seqFISH, respectively, where both graphs present significant author clusters. Similarly, Figure 2C provides the projection graph of CUGBP—a regulatory factor that has not been extensively studied. Of note, in the small projections, we labeled the author collaborations, i.e. edges connecting two authors. The collaboration subgraph is the same as that in Extended Data Supplementary Figure S8, in which the citation counts in Google Scholar of their multi-omics papers determine the node weight (see methods in Section 'Mapping of bipartite author-topic graph and its projections' for details). Tight connections between authors within the same research group and across groups can be observed in the graph. Researchers with prominent reputations in the field (e.g., inventor of the selected imaging techniques in the projections of Figure 2B and D) tend to have larger sizes and more connections, and thus have larger degrees.

With the ability to filter for a specific topic or multiple entities (e.g. different FISH methods or derivatives, or a set of author names, proteins or genes input by the user), we anticipate the ability of iSMOD to integrate data or experimental results from a sub-network to draw biological conclusions or hypotheses.

Investigating the molecular mechanism of diseases by combining the FISH results of the genes or alleles within a given chromatin region. In iSMOD, there are currently 10 848 and 7541 genes from human (Figure 3A) and mouse (Figure 3B) chromosomes, respectively, imaged using FISH. An important application of mining the enormous data is to retrieve all FISH articles referring to the genes within the chromosome region associated with a disease and infer or validate the mechanism underlying the disease. For example, in human chromosomes, for the genome region `Chr11:5225464--5269945` containing the gene `HBB` that encodes β-globin (a subunit of the important protein hemoglobin that is present within red blood cells and binds to oxygen molecules in the lungs), eight FISH articles can be found in iSMOD. This set of papers potentially refers to the related studies revealing the mechanisms and functions of this chromosomal region. Miles *et al.* reported (42) that the intergenic transcription of a 20-kb sub-domain in this region occurs outside the S phase, and a high level of active modification is obtained, primarily for histones `H3`. As an important counterpart of histone `H3`, the enrichment of the histone modification `H3K9me3` has been found to support the peripheral localization of chromatin regions called `lamina-associated domains (LADs)`, where the peripheral targeting of the β-globin gene cluster, including `HBB`, serves as a representative, and the cluster targets the nuclear interior during erythrocyte maturation while transcription occurs (43). Regarding the cause of the up-regulation of `HBB` and four other genes (`HBA`, `SLC4A1`, `ERAF` and `GATA1`) that occurs during erythropoiesis, Brown *et al.* (44) reported significant but varying association levels among the five genes (which are located on four different chromosomes), and found that M-FISH further validated the proximity of `HBA`, `HBB`, `SLC4A1`, and `ERAF`. These results are primarily attributed to the effects of splicing factor aggregations within the nuclear speckle, as summarized in Figure 3C. In addition to revealing the among-gene associations with blood oxygen transportation, the other retrieved papers can provide observation and drug mechanism insights related to β-globin. For example, Stavrou *et al.* (45) reported that the β-globin replicator produces vectors that promote transfection efficiency in hematopoietic progenitor cells, in which the regulation can be observed with enhanced green fluorescent protein-tagged plasmids. Moreover, CRISPR/Cas9-mediated knock-in can localize `HBB` without perturbation and demonstrate the long-term stability of `HBB` (46). Finally, as a regulated gene, `HBB` plays an important role in drug mechanism research to verify specific toxicity (47).

Monoallelic expression due to genomic imprinting or allelic exclusion can be visualized using allele-specific FISH, which offers to determine the abnormality of single alleles in chromosomes and investigate the mechanisms involved in genomic imprinting. iSMOD permits users to search allele-specific FISH articles with >2000 allele-specific papers. Among them, a total of 118 alleles imaged using FISH in human chromosomes, and 109 alleles imaged using FISH in mouse chromosomes have been identified; their distributions are shown in Figure 3D and Figure 3E, respectively.

For example, the Beckwith–Wiedemann (`BWS OMIM #130650`) and Silver–Russell (`SRS OMIM #180860`) syndromes exhibit opposite growth abnormalities. Collected evidence from allele-specific FISH and other experiments shows that both syndromes are caused by (epi)genetic defects at `11p15.5` (`Chr11:1-2 800 000`) (48). For this 1-Mb-long imprinted region, 39 papers were retrieved from iSMOD, we then rapidly identified the FISH experiments conducted for `H19` (imprinted maternally expressed

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkad580/7223586 by guest on 04 August 2023

transcript) and Igf2 (insulin-like growth factor 2) (49,50) to improve our understanding of the expression pattern of the H19-Igf2 imprinting gene cluster. Eggermann *et al.* (48) revealed that on unmethylated maternal allele, CTCF binding results in a boundary to prevent Igf2 promoters from accessing enhancers, and conversely, methylation of the differentially methylated region (DMR) on the paternal allele prevents CTCF binding. Furthermore, according to Rovina *et al.* (49), in addition to the differential methylation of DMR within the H19/IGF2 domain, the interaction between the H19/IGF2 and CDKN1C/KCNQ1OT1 domains also has a strong impact on the pathogenesis of these two syndromes. In the CDKN1C/KCNQ1OT1 domain, KCNQ1OT1 is an imprinted antisense long non-coding RNA (lncRNA) on chromosome 11p15.5 which participates in cell proliferation, migration (51), and chromosomal domain localization (52); this is considered to regulate the development of colorectal cancer (53) and diseases such as retinal infection (54). Refined queries in iSMOD reveal the imprinting mechanism of CDKN1C/KCNQ1OT1, which is similar to that of H19/IGF2 in normal cases, and the Kcnq1ot1 imprinting control region (ICR) is unmethylated on the paternal allele, which allows Kcnq1ot1 ncRNA transcription (55). The expression level is regulated by various nucleoporins (55) or β-catenin (53) under different circumstances, and this process acts as a microRNA sponge to regulate the expression of other RNAs (51). Specifically, the reciprocal expression of these two regions is regulated in a methylation-sensitive manner by competing with a shared set of enhancers. There are three interactions in this region—between imprinting control region 1(ICR1) and enhancer candidate regions located upstream of ICR2 (Enh 2), between ICR2 and the region upstream of IGF2, and between ICR2 and CTCF Dw—among which, the ICR2--CTCF Dw association is inferred to occur on the maternal allele, whereas the ICR2--upstream IGF2 interaction may occur on the paternal allele. Other putative enhancer regions have also been suggested to potentially affect CDKN1C expression mediated by CTCF (56). In pathological conditions, the ICR1--Enh 2 and ICR2--CTCF Dw interactions are lost, and new interactions, such as ICR2--Enh A and ICR2--Enh B, appear. Research on this chromosomal range also involves deletions/duplications of imprinted loci (57), which are often associated with the mechanisms of imprinting control. As validated by user studies in one retrieved paper (58), an *in cis* duplication of the entire 11p15.5 cluster on the maternal allele may be related to the disease phenotype. Furthermore, the microduplication of truncated KCNQ1OT1 may lead to the presence of both methylated and unmethylated ICR2 sequences on the same chromosome, resulting in CDKN1C silencing on the maternal chromosome but causing no effect on the paternal chromosome. Upon combining the results from the retrieved articles, a conceptual map was constructed, as shown in Figure 3F, which illustrates the regulatory mechanisms and chromosomal transitions underlying the Beckwith–Wiedemann and Silver–Russell syndromes.

As demonstrated above, the integration of research conclusions at different levels provides comprehensive insights into the interactions and activities within and between gene regions. We anticipate that with an increase in published articles, a more holistic understanding of various genomic and molecular mechanisms underlying diseases can be elucidated with the aid of iSMOD and data reanalysis. In iSMOD, we have annotations for 131 cancer types and 642 cancer cell lines (Figure 3G–H). One can search for related papers for a specific cancer type using combinatorial analysis and uncover the underlying mechanisms to further understand how the genetic changes in 3D result in key genetic variants that can ultimately lead to cancer development.

Exploring the working mechanisms under biological phenomena by mining multi-omics interactions from the iSMOD repository. The entity–entity connection graph refers to the interaction relationships established from semantic analysis of the interactions among molecular entities (e.g. promoters, enhancers, genes, alleles, transcription factors, etc.), representing each entity as a node and interactions as edges. Together with customized searching, iSMOD can build a graph based on retrieved papers on pairwise interactions related to the topic of interest at varying omics levels and integrate the results to reveal a portion of the underlying working mechanisms. Here, we use liquid–liquid phase separation (LLPS) as a representative, which is increasingly being used as a potential physicochemical basis for the formation of membrane-less bodies in cells, such as in promyelocytic leukemia nuclear body (PML NB) (59). Since LLPS is closely related to various nuclear proteins and plays a key role in regulating the relationship among proteins, RNA, and chromatin, integrating the locations (determined by FISH) of and interactions among these entities may help reveal the mechanisms underlying phase separation. iSMOD produced a graph using the data from 82 retrieved related papers and presented it as filled nodes together with their related edges (Figure 4A). We supplemented the graph manually with hollow nodes, as a limited number of entities were not retrieved in several papers. The percentage of the filled nodes largely validates the effectiveness of the automatic extraction of interactions. From the graph, we can draw the following conclusions:

(i) iSMOD helps mine the molecular connections underlying separated studies on LLPS of various proteins or complexities, for example, the role of the intrinsically disordered region (IDR) illustrated by the 'PML NB' cluster in Figure 4A. Experimentally, the existence of LLPS condensate formed by BRD4, MED1, and RNAPII indicates that proteins with a higher content of disordered regions are associated with higher LLPS potential (60), revealing a close connection between IDR and LLPS. The alternative lengthening of telomeres associated with the condensation of PML NB (APB) is also driven by interactions between the small ubiquitin-like modifier (SUMO) and the SUMO interaction motif used for eDHFR (61). SUMO1 modification enables LLPS and is primarily present within the outer shell of PML NBs. Modification plays a key role in the recruitment and partitioning of regular 'client' proteins with different functions, for example, transcriptional regulation (HDAC7), heterochromatin establishment (SETDB1), and chromatin dynamics regulation

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkad580/7223586 by guest on 04 August 2023

(SIRT1) (59). Furthermore, besides the modification of SUMO, the IDR of PML is also essential for LLPS, as demonstrated by the fact that disordered C-terminal domains of PML isoforms tend to cause the initiation of LLPS (62), whereas those containing no disordered structures are incapable of LLPS (63).
Strong connections between IDR and LLPS have also been reported in coronavirus disease 2019 (COVID-19) studies. Investigations into severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) suggest that the multifunctional nucleocapsid (N) protein in its central IDR drives phase separation with RNA (64), which is caused by interactions between R2 motifs (amino acids 369–390) (65). FISH experiments on rotavirus at the early stages of infection further demonstrate that the viroplasm is formed by interactions between the RNA chaperone NSP2 and the intrinsically disordered protein NSP5 and its co-localization with RNA (66).

(ii) Phase separation plays an important role in DNA damage repair (DDR), as illustrated by the 'DDR factors' cluster in Figure 4A. FUS is an RNA-binding protein (RBP) that participates in DDR at sites of KU80, NBS1, and SFPQ (67). The LLPS of FUS can be caused by amyotrophic lateral sclerosis (ALS)-linked mutations. The LLPS mediated by FUS protein is modulated by specific (NEAT1 RNA) and non-specific RNAs (68). The LLPS of FUS recruits LAMP1-positive structures (69). As an important requirement for the recruitment of key proteins, such as XRCC1, to DNA damage sites (70), LLPS is essential for the proper formation of DNA damage foci and activation of the DDR signaling cascade, thus serving as the factor necessary for the participation of FUS in DDR (67).

(iii) The ribonucleoprotein (RNP) milieu is formed by the interaction of lncRNA and IDRs in RBPs, where the phase separation is established under the participation of HRSs (71). Combining the two retrieved papers (72,73), we sketched the generation of the RNP milieu, as shown in the 'HRS-related genes' cluster and its neighboring nodes in Figure 4A: As an example of the lncRNA–protein complex, paraspeckle is formed by the phase separation of associating RBPs, where lncRNAs transcribed by enhancer or promoter regions are thought to form a non-membranous RNP milieu through association with IDRs, in which a key step is the multimerization of SFPQ and NONO in the middle region of NEAT1_2,nakagawa2018molecular (74). Large RNP complexes also exhibit deep connections with high-salt–recovered sequences (HRSs), i.e. the genomic DNA associated with insoluble materials, such as coilin, SMN, and PML obtained after high-salt treatment on genes. These sequences correspond to Sox2, Pou5f1, Nanog and Klf4 loci, which may be associated with a large RNP complex in cortical neurons and map to Malat1 / Neat1 loci as well (73), which are necessary for the assembly of nuclear speckles/paraspeckles and the regulation of gene expression (75).

(iv) The dynamics of nuclear speckle LLPS are regulated by a 12-h ultradian rhythm established by the XBP1s–SON axis, as shown in the 'XBP1s-SON axis' cluster in Figure 4A. iSMOD retrieved a set of interactions from the paper of Dion *et al.* (76) and their connections to other nodes. A simple analysis of the related text describes the following pathway: Immunofluorescence against the SRSF2(SC35) marker confirms that the LLPS dynamics (nuclear speckle morphology) are regulated by XBP1 and its downstream gene SON with a 12-h ultradian rhythm, and the Neat1–Malat1 binding serves as an illustration of the nuclear speckle–chromatin interactions, validating the regulation of XBP1. At the genomic level, sets of genes, such as Manf, Hyou1 and Sec23b were found to be hypersensitive to the dynamic changes of LLPS (76).

Overall, benefiting from the decent extraction of interactions, iSMOD can potentially reveal a wide range of phenomena, mechanisms, and associations for various interactions among molecular entities.

Integration of 3D genomic, transcriptomic, and proteomic data in a virtual cell nucleus. Integrating the locations of the same gene at different time points from different articles into a unified coordinate system is beneficial for the inference of its migration and spreading mechanism during cell development. Such integration can form a 'virtual' cell that can function as the gold standard to verify the gene locations in other experiments, including those comprising ligation-based (Hi-C (14), capture-C (77)), or ligation-free methods (GAM (78) and SPRITE (79)). Moreover, for a specific cell type from the same species, at a given phase in the cell cycle, all genes with FISH results can be integrated into a unified 3D coordinate system to obtain a complete spatial distribution of the genes. For example, integrating the coordinates of genes determined using seqFISH with those from other techniques yields a more comprehensive 3D map of multiple genes. Although more than 14,500 FISH papers are collected in iSMOD, the FISH data for the same gene from different papers are not yet sufficient to generate a complete virtual cell. Nevertheless, there are promising examples that can be used to validate the importance of iSMOD and advance efforts in single-cell multi-omics data sharing.

For instance, X inactive specific transcript (Xist), an RNA gene, acts as a major effector of the X-chromosome-inactivation process. We can register and merge the data from two papers (33,34) to obtain a more complete illustration (see Figure 4B). For the SeqFISH data by Takei *et al.* (33), we quantitated the 3D position of Xist using shells of equal radius division (see Figure 4C) and found that Xist is distributed randomly in the nucleus (*t*-test: $P = 0.65$; Mann–Whitney *U*-test: $P = 0.82$; one-way ANOVA: $P = 0.65$; Figure 4D). By contrast, the Xist data by Shiura *et al.* (34) are not randomly distributed (Extended Data Supplementary Table S2) but located more frequently in the central region between the nuclear periphery and nuclear geometric center. This finding is primarily attributed to the fact that cells in the two papers are in different phases. Such integration can be used to visualize the differences in status and validate the impacts of specific loci in different phases. For example, in their movies on the spreading process of

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkad580/7223586 by guest on 04 August 2023

Xist RNA, Namekawa *et al.* (80) reveal the importance of Xist in gene silencing at the morula stage (the second step of imprinted X inactivation). iSMOD can thus make it easier to reveal such information by exhibiting the spatial distributions of Xist at different developmental phases by integrating partial data from different groups into the proposed virtual cell.

We can also set up a global 3D distribution map to visualize the available transcriptomic data for the same cell type, from the public MERFISH data released by Vizgen Data Release Program (https://info.vizgen.com/mouse-brain-data) and other related papers in iSMOD.

First, we demonstrated the ability of iSMOD to integrate RNA-seq data, where the transcripts in the brain cells imaged using MERFISH were visualized in 3D space (Figure 4E). Furthermore, by considering multi-omics data such as those for nucleoli and speckles, another 'virtual cell' was built using the MERFISH data from IMR-90 cells (35) by integrating the spatial information for genome-scale chromatin organization, transcriptional activity, and nuclear structures into one 3D map (Figure 4F). The zoomed-in model shown in Figure 4G demonstrates the detailed distributions of multi-omics entities; the genes that overlapped with homeobox (HOX) gene family on chromosomes are marked in purple, with the potential for further exploration. All virtual cells shown in Figure 4 are modeled on the homepage of iSMOD in an interactive manner (see methods in Section 'The construction of 3D virtual cells' for details). This demonstrates that iSMOD is a good candidate for the single-cell multi-omics data platform to investigate the mechanisms of gene regulation.

SUMMARY AND DISCUSSION

This study aimed to develop a method for advanced biological studies that unify multi-omics data and technologies. Consequently, the image-based multi-omics search engine iSMOD was developed, using newly collected and annotated data for 23 288 (which continues to grow) life sciences and medicine papers from PubMed. A new website, https://i-smod.com, was constructed to search, browse, and analyze the papers included in the database with a user-friendly GUI. A salient feature of the database is that it currently includes at least 11,664 genes/genomic loci or RNAs and 142 proteins imaged using DNA FISH, RNA FISH and immune-staining, respectively.

As the first integrated single-cell multi-omics platform, iSMOD provides a new means to serve the life sciences and medical community in an interdisciplinary manner. We have detailed typical examples to give a broader perspective of the possibilities the tool is capable of: (i) It is envisioned that the platform will inspire researchers to conduct deeper data mining to combine the fragmentary conclusions scattered in tens of thousands of papers and investigate the spatiotemporal events underlying diseases, such as tumorigenesis and various rare diseases (48,51,57). (ii) This search engine will enable researchers to visualize the nuclear structure and molecular activities in a unified coordinate system, thus advancing the current understanding regarding the relationships among chromatin organization, gene regulation, transcription, and translation. Such a unified view will largely benefit from recent progress in high-throughput and multimodal imaging based on multiplexed FISH imaging methods (e.g. MERFISH and seqFISH) and directly links genomic loci with nascent RNA transcripts and landmark nuclear structures. (iii) iSMOD serves as a platform to identify the key studies and researchers involved in the multi-omics fields, which will help facilitate a more rapid understanding of the history, key technologies, hot topics, productive research groups, and author collaborations. This profiling is efficient and informative as the results are built on comprehensive annotations from the contents of academic papers rather than keywords and titles. (iv) The rich set of annotations and search options make iSMOD a powerful tool that can meet the demands of different users. For example, the function of retrieving papers related to a specific gene or genes within a chromosomal region is useful for studies on a target topic; searching the papers and experimental results of a given dye can quickly identify the proper dyes from previous investigations; one can also retrieve the entries related to a specific cancer type or cell line.

iSMOD also provides ranking to the search results to help users identify the expected literature. For example, in papers where diseases or genes are mentioned in the abstract or title but not related to FISH or protein analysis, the annotations of these papers tend to provide misleading results. However, articles, especially meta-analyses and review articles, containing high citation count or published in high-impact journals tend to be more rigorous in the textual description of the title and abstract, thus serving to potentially provide more relevant information. To this end, iSMOD provides ranking by either publication time or citation count, helping users to retrieve the relevant articles efficiently.

FUTURE DIRECTIONS

As multi-omics techniques are developing and evolving rapidly, related academic papers are being published at an increasing rate, and the demand for systematic reviewing and integration of related studies is becoming even more important. Correspondingly, an increasing amount of image-based multi-omics data will be published, and this will provide large amounts of new data relevant to iSMOD . To better track the rapid progress in this field and to benefit the academic community, we will continue to update iSMOD with the latest published work to help reveal promising research trends.

We also plan to expand the platform to integrate image data from other omics, including image-based transcriptomic data and single nuclear metabolomic data, to serve the integrative single-cell multi-omics studies in a more comprehensive manner. It is also worth noting that the research on disease and clinical symptoms accounts for a large percentage of the articles in iSMOD , where quantitative data are provided and can potentially reveal the pathogenesis of disease statistically. Therefore, iSMOD is further expected to incorporate the results of Kaplan—Meier analysis as well as genotype (mutations) and phenotype information that is crucial for depicting the research conclusion, if provided, in the search results in the future. Moreover, integrating spatial multi-omics data or the results obtained

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkad580/7223586 by guest on 04 August 2023

in a multi-cellular system can advance the investigations of cell–cell and cell–extracellular matrix interactions in a similar manner. In the future, we plan to construct a browser for spatial multi-omics data, extending the database of integrative biomedical studies to a larger scale, where studies ranging from molecular analyses to cell arrangements and regulation would be included.

Servers for multi-omics datasets and analysis have become booming tools in recent years, and some representatives in these works introducing similar organizing servers for experimental/computational data are important references (81–83) and data sources (25,26,84) for iSMOD. However, manuscripts that primarily organize existing experimental/computational data may probably not enter our database due to lacks of conclusive experimental images according to the information extraction pipeline. To this end, we consider provide database access to multi-omics experimental/computational data in the "Helpful Links" column of iSMOD, and one can retrieve/locate such servers using search engines with proper keywords.

While advancing toward more comprehensive search results, achieving the pipeline from the output of iSMOD to reliable conclusions or predictions serves as another interesting direction. Currently, it does require a certain amount of professional knowledge to interpret the aggregated results or retrieve meaningful connections from the interaction graph. To this end, the large language models (85)/general foundation models (arXiv preprint: 2108.07258) have demonstrated their potential in analyzing natural languages and extracting the main conclusions in the article, which further becomes a requisite for providing suggestions for unknown mechanisms/pathways. Moreover, in-context learning is widely explored and making progress in large language models (arXiv preprint arXiv:2202.12837, arXiv preprint arXiv:2301.00234). Therefore, a possible solution for the conclusion-drawing function is expected to conduct transfer learning (86) on the foundation large language model by using the text from multi-omics research, and the fine-tuned model could provide effective output as another important annotation of iSMOD, where large-scale data collection, environmental configuration, and hardware upgrades are required. Therefore, subsequent versions of iSMOD will focus on the latest developments in foundation model research and move toward integration with high-level semantic analysis tools. Moreover, there are a lot of computational works utilizing image-based omics data, which plays a key part in multi-omics studies. The computational works containing experimental figures or panels will be labeled by the keywords and enter our database, assisting researchers to integrate the conclusions computationally drawn from image-based omics data. In the future, utilizing advanced semantic analysis techniques for gathering omics data sources, recognizing the computation algorithm, and retrieving the computational conclusion from such computational works, we can combine experimental and computational works for more comprehensive information integration. Besides, based on the gathered omics data, causal learning method (87) that has already demonstrated exceptional results for climate prediction and cardiovascular monitoring (88), has immense potential to explore correlations between multi-omics data and phenotypes in research papers.

Additionally, a more comprehensive annotation is expected to provide the interaction graph with advanced filtering and retrieving functions, thereby increasing the ease of exploration.

While iSMOD is a powerful search engine, data integration remains limited, as most 3D FISH or protein data were identified in the original papers, however, were inaccessible due to various reasons, including hard disk clashes or obsolete corresponding emails. We have contacted the corresponding authors of over 100 retrieved papers but received only seven positive responses (34,89–94) to date. Thus, limited data availability has become one potential challenge for the real-time modeling of virtual cells based on search results. In addition, although it is possible to collect the source data information (if any) in most articles through automated programs, most source data provided in omics papers are very large and demand extended processing time and large-scale storage occupation, which is not affordable for iSMOD's server. Furthermore, data identification and pre-processing often rely on empirical and manual operations. Therefore, generating virtual cells online based on real-time search results on the website is currently infeasible but may push the process one step forward by extracting the corresponding information provided by iSMOD. In the future, we plan to collaborate as a data-sharing site to help maintain valuable data for future integrative studies. Toward more versatile integration for virtual cells, iSMOD will continue to encourage wider data availability and promote efficient processing combined with advanced computer vision algorithms.

Another issue to be reckoned with is that aggregating the 3D coordinate information solely based on specific cell types or experimental methods may result in decreased accuracy due to factors such as differentiated experimental conditions. In the future, we plan to deploy advanced semantic models for comprehensive annotations on experimental conditions/controls/benchmarks, and a compound analysis that includes histology images and gene expression abundance to facilitate the constructions of refined and accurate 3D spatiomics models, where the graph convolutional network has been proposed to identify spatial domains (e.g., the 3D expression domain in MERFISH data) and has demonstrated good scalability (95,96); however, certain issues must first be resolved to allow its implementation in iSMOD: First, retrieving and annotating histological images from the article is required based on the establishment of a spatial multi-omics data browser. Second, better data availability is encouraged to match the histology images as the compound input of the deep learning model. Third, an updated neural network is necessary to generalize various omics images and data. In summary, achieving more accurate integration not only poses challenges for the future version of iSMOD construction, but is also closely related to the advance of image processing, deep learning, and multi-omics research itself.

Furthermore, as is common with other search engines, not all extracted keywords are closely related to the topic of the article and not all the articles are directly related to the FISH or proteins research, although various approaches

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkad580/7223586 by guest on 04 August 2023

described in Section 'iSMOD data collection and annotation' have been adopted. In the future, rapidly improving semantic analysis algorithms for the article and search items will be applied to produce better annotation and more accurate results. We also plan to extend the search range in the main text and include annotations on the supplementary materials.

Although most steps of iSMOD construction are automatic, the acquisition of citation counts of articles and the creation of knowledge graphs requires some manual assistance; thus, iSMOD is not updated in real time. Instead, considering the rapid advances in the field of single-cell multi-omics, iSMOD will maintain a bi-monthly regular update. The GUI and functions will continue to be improved or supplemented based on the received feedback from researchers.

DATA AVAILABILITY

No new data were generated or analyzed in support of this research.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Prof. Minping Qian from Peking University for the helpful discussions.

FUNDING

Beijing Municipal Natural Science Foundation [Z200021 to J.S. and J.G., in part]; National Natural Science Foundation of China [81890991]; State Key Research Development Program of China [2017YFA0505503, 2021YFE0201100]; CAS Interdisciplinary Innovation Team [JCTD-2020-04 to J.G.]. Funding for open access charge: Beijing Municipal Natural Science Foundation [Z200021].
Conflict of interest statement. None declared.

REFERENCES

1. Bourne,P.E., Lorsch,J.R. and Green,E.D. (2015) Perspective: sustaining the big-data ecosystem. *Nature*, **527**, S16–S17.
2. Prins,P., De Ligt,J., Tarasov,A., Jansen,R.C., Cuppen,E. and Bourne,P.E. (2015) Toward effective software solutions for big biology. *Nat. Biotech.*, **33**, 686–687.
3. Eisenstein,M. (2022) Seven technologies to watch in 2022. *Nature*, **601**, 658–661.
4. Lichter,P., Cremer,T., Borden,J., Manuelidis,L. and Ward,D. (1988) Delineation of individual human chromosomes in metaphase and interphase cells by in situ suppression hybridization using recombinant DNA libraries. *Hum. Genet.*, **80**, 224–234.
5. O'Connor,C. (2008) Fluorescence in situ hybridization (FISH). *Nat. Educ.*, **1**, 171.
6. Chen,K.H., Boettiger,A.N., Moffitt,J.R., Wang,S. and Zhuang,X. (2015) Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, **348**, aaa6090.
7. Lubeck,E., Coskun,A.F., Zhiyentayev,T., Ahmad,M. and Cai,L. (2014) Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods*, **11**, 360–361.
8. Shah,S., Lubeck,E., Zhou,W. and Cai,L. (2016) In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron*, **92**, 342–357.
9. Codeluppi,S., Borm,L.E., Zeisel,A., La Manno,G., van Lunteren,J.A., Svensson,C.I. and Linnarsson,S. (2018) Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat. Methods*, **15**, 932–935.
10. Goh,J. J.L., Chou,N., Seow,W.Y., Ha,N., Cheng,C. P.P., Chang,Y.-C., Zhao,Z.W. and Chen,K.H. (2020) Highly specific multiplexed RNA imaging in tissues with split-FISH. *Nat. Methods*, **17**, 689–693.
11. Liu,M., Yang,B., Hu,M., Radda,J.S., Chen,Y., Jin,S., Cheng,Y. and Wang,S. (2021) Chromatin tracing and multiplexed imaging of nucleome architectures (MINA) and RNAs in single mammalian cells and tissue. *Nat. Protoc.*, **16**, 2667–2697.
12. Mateo,L.J., Murphy,S.E., Hafner,A., Cinquini,I.S., Walker,C.A. and Boettiger,A.N. (2019) Visualizing DNA folding and RNA in embryos at single-cell resolution. *Nature*, **568**, 49–54.
13. Shaffer,S.M., Dunagin,M.C., Torborg,S.R., Torre,E.A., Emert,B., Krepler,C., Beqiri,M., Sproesser,K., Brafford,P.A., Xiao,M. *et al.* (2017) Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature*, **546**, 431–435.
14. Lieberman-Aiden,E., Van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
15. Fullwood,M.J., Liu,M.H., Pan,Y.F., Liu,J., Xu,H., Mohamed,Y.B., Orlov,Y.L., Velkov,S., Ho,A., Mei,P.H. *et al.* (2009) An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, **462**, 58–64.
16. Shin,Y., Chang,Y.-C., Lee,D.S., Berry,J., Sanders,D.W., Ronceray,P., Wingreen,N.S., Haataja,M. and Brangwynne,C.P. (2019) Liquid nuclear condensates mechanically sense and restructure the genome. *Cell*, **176**, 1518.
17. Shin,Y. and Brangwynne,C.P. (2017) Liquid phase condensation in cell physiology and disease. *Science*, **357**, eaaf4382.
18. Vizcaíno,J.A., Deutsch,E.W., Wang,R., Csordas,A., Reisinger,F., Ríos,D., Dianes,J.A., Sun,Z., Farrah,T., Bandeira,N. *et al.* (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotech.*, **32**, 223–226.
19. Samaras,P., Schmidt,T., Frejno,M., Gessulat,S., Reinecke,M., Jarzab,A., Zecha,J., Mergner,J., Giansanti,P., Ehrlich,H.-C. *et al.* (2020) ProteomicsDB: a multi-omics and multi-organism resource for life science research. *Nucleic Acids Res.*, **48**, D1153–D1163.
20. Desiere,F., Deutsch,E.W., King,N.L., Nesvizhskii,A.I., Mallick,P., Eng,J., Chen,S., Eddes,J., Loevenich,S.N. and Aebersold,R. (2006) The peptideatlas project. *Nucleic Acids Res.*, **34**, D655–D658.
21. Perez-Riverol,Y., Bai,M., da Veiga Leprevost,F., Squizzato,S., Park,Y.M., Haug,K., Carroll,A.J., Spalding,D., Paschall,J., Wang,M. *et al.* (2017) Discovering and linking public omics data sets using the Omics Discovery Index. *Nat. Biotech.*, **35**, 406–409.
22. Durand,N.C., Robinson,J.T., Shamim,M.S., Machol,I., Mesirov,J.P., Lander,E.S. and Aiden,E.L. (2016) Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.*, **3**, 99–101.
23. Kerpedjiev,P., Abdennur,N., Lekschas,F., McCallum,C., Dinkla,K., Strobelt,H., Luber,J.M., Ouellette,S.B., Azhir,A., Kumar,N. *et al.* (2018) HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biol.*, **19**, 125.
24. Robinson,J.T., Turner,D., Durand,N.C., Thorvaldsdóttir,H., Mesirov,J.P. and Aiden,E.L. (2018) Juicebox. js provides a cloud-based visualization system for Hi-C data. *Cell Syst.*, **6**, 256–258.
25. Cunningham,F., Allen,J.E., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Austine-Orimoloye,O., Azov,A.G., Barnes,I., Bennett,R. *et al.* (2022) Ensembl 2022. *Nucleic Acids Res.*, **50**, D988–D995.
26. Martin,F.J., Amode,M.R., Aneja,A., Austine-Orimoloye,O., Azov,A.G., Barnes,I., Becker,A., Bennett,R., Berry,A., Bhai,J. *et al.* (2023) Ensembl 2023. *Nucleic Acids Res.*, **51**, D933–D941.
27. Chang,F.T., McGhie,J.D., Chan,F.L., Tang,M.C., Anderson,M.A., Mann,J.R., Andy Choo,K. and Wong,L.H. (2013) PML bodies provide an important platform for the maintenance of telomeric chromatin integrity in embryonic stem cells. *Nucleic Acids Res.*, **41**, 4447–4458.
28. Guimerà,R. and Nunes Amaral,L.A. (2005) Functional cartography of complex metabolic networks. *Nature*, **433**, 895–900.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkad580/7223586 by guest on 04 August 2023

29. Barabási,A.-L., Jeong,H., Néda,Z., Ravasz,E., Schubert,A. and Vicsek,T. (2002) Evolution of the social network of scientific collaborations. *Phys. A: Stat. Mech. Appl.*, **311**, 590–614.
30. Watts,D.J. and Strogatz,S.H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–442.
31. Otte,E. and Rousseau,R. (2002) Social network analysis: a powerful strategy, also for the information sciences. *J. Inform. Sci.*, **28**, 441–453.
32. Floyd,R.W. (1962) Algorithm 97: shortest path. *Commun. ACM*, **5**, 345.
33. Takei,Y., Yun,J., Zheng,S., Ollikainen,N., Pierson,N., White,J., Shah,S., Thomassie,J., Suo,S., Eng,C.-H.L. *et al.* (2021) Integrated spatial genomics reveals global architecture of single nuclei. *Nature*, **590**, 344–350.
34. Shiura,H. and Abe,K. (2019) Xist/Tsix expression dynamics during mouse peri-implantation development revealed by whole-mount 3D RNA-FISH. *Sci. Rep.*, **9**, 3637.
35. Su,J.-H., Zheng,P., Kinrot,S.S., Bintu,B. and Zhuang,X. (2020) Genome-scale imaging of the 3D organization and transcriptional activity of chromatin. *Cell*, **182**, 1641–1659.
36. Katz,J.S. and Martin,B.R. (1997) What is research collaboration?. *Research Policy*, **26**, 1–18.
37. Price,D.J., Merton,R.K. and Garfield,E. (1986) In: *Little Science, Big Science... and Beyond*. Columbia University Press, NY, Vol. **480**.
38. Zuccala,A. (2006) Modeling the invisible college. *J. Am. Soc. Inform. Sci. Technol.*, **57**, 152–168.
39. Wasserman,S. and Faust,K. (1994) Social network analysis: methods and applications. In *Structural Analysis in the Social Sciences*. Cambridge University Press.
40. Freeman,L. (2004) The development of social network analysis. *A Study in the Sociology of Science*, **1**, 159–167.
41. Freeman,L.C. (1977) A set of measures of centrality based on betweenness. *Sociometry*, 35–41.
42. Miles,J., Mitchell,J.A., Chakalova,L., Goyenechea,B., Osborne,C.S., O'Neill,L., Tanimoto,K., Engel,J.D. and Fraser,P. (2007) Intergenic transcription, cell-cycle and the developmentally regulated epigenetic profile of the human beta-globin locus. *PLoS One*, **2**, e630.
43. Bian,Q., Khanna,N., Alvikas,J. and Belmont,A.S. (2013) β-Globin cis-elements determine differential nuclear targeting through epigenetic modifications. *J. Cell Biol.*, **203**, 767–783.
44. Brown,J.M., Green,J., das Neves,R.P., Wallace,H.A., Smith,A.J., Hughes,J., Gray,N., Taylor,S., Wood,W.G., Higgs,D.R. *et al.* (2008) Association between active genes occurs at nuclear speckles and is modulated by chromatin environment. *J. Cell Biol.*, **182**, 1083–1097.
45. Stavrou,E.F., Lazaris,V.M., Giannakopoulos,A., Papapetrou,E., Spyridonidis,A., Zoumbos,N.C., Gkountis,A. and Athanassiadou,A. (2017) The β-globin Replicator greatly enhances the potential of S/MAR based episomal vectors for gene transfer into human haematopoietic progenitor cells. *Sci. Rep.*, **7**, 40673.
46. Tasan,I., Sustackova,G., Zhang,L., Kim,J., Sivaguru,M., HamediRad,M., Wang,Y., Genova,J., Ma,J., Belmont,A.S. *et al.* (2018) CRISPR/Cas9-mediated knock-in of an optimized TetO repeat for live cell imaging of endogenous loci. *Nucleic Acids Res.*, **46**, e100.
47. Alsagaby,S.A., Vijayakumar,R., Premanathan,M., Mickymaray,S., Alturaiki,W., Al-Baradie,R.S., AlGhamdi,S., Aziz,M.A., Alhumaydhi,F.A., Alzahrani,F.A. *et al.* (2020) Transcriptomics-based characterization of the toxicity of ZnO nanoparticles against chronic myeloid leukemia cells. *Int. J. Nanomed.*, **15**, 7901.
48. Eggermann,T., Eggermann,K. and Schönherr,N. (2008) Growth retardation versus overgrowth: Silver-Russell syndrome is genetically opposite to Beckwith-Wiedemann syndrome. *Trends Genet.*, **24**, 195–204.
49. Rovina,D., La Vecchia,M., Cortesi,A., Fontana,L., Pesant,M., Maitz,S., Tabano,S., Bodega,B., Miozzo,M. and Sirchia,S.M. (2020) Profound alterations of the chromatin architecture at chromosome 11p15. 5 in cells from Beckwith-Wiedemann and Silver-Russell syndromes patients. *Sci. Rep.*, **10**, 8275.
50. Fazi,B., Garbo,S., Toschi,N., Mangiola,A., Lombari,M., Sicari,D., Battistelli,C., Galardi,S., Michienzi,A., Trevisi,G. *et al.* (2018) The lncRNA H19 positively affects the tumorigenic properties of glioblastoma cells and contributes to NKD1 repression through the recruitment of EZH2 on its promoter. *Oncotarget*, **9**, 15512.
51. Wang,J., Zhang,H., Situ,J., Li,M. and Sun,H. (2019) KCNQ1OT1 aggravates cell proliferation and migration in bladder cancer through modulating miR-145-5p/PCBP2 axis. *Cancer Cell Int.*, **19**, 325.
52. Fedoriw,A.M., Calabrese,J.M., Mu,W., Yee,D. and Magnuson,T. (2012) Differentiation-driven nucleolar association of the mouse imprinted Kcnq1 locus. *G3: Genes| Genomes| Genetics*, **2**, 1521–1528.
53. Sunamura,N., Ohira,T., Kataoka,M., Inaoka,D., Tanabe,H., Nakayama,Y., Oshimura,M. and Kugoh,H. (2016) Regulation of functional KCNQ1OT1 lncRNA by β-catenin. *Sci. Rep.*, **6**, 20690.
54. Rochet,E., Appukuttan,B., Ma,Y., Ashander,L.M. and Smith,J.R. (2019) Expression of long non-coding RNAs by human retinal müller glial cells infected with clonal and exotic virulent toxoplasma gondii. *Non-coding RNA*, **5**, 48.
55. Sachani,S.S., Landschoot,L.S., Zhang,L., White,C.R., MacDonald,W.A., Golding,M.C. and Mann,M.R. (2018) Nucleoporin 107, 62 and 153 mediate Kcnq1ot1 imprinted domain regulation in extraembryonic endoderm stem cells. *Nat. Commun.*, **9**, 2795.
56. López-Abad,M., Iglesias-Platas,I. and Monk,D. (2016) Epigenetic characterization of CDKN1C in placenta samples from non-syndromic intrauterine growth restriction. *Front. Genet.*, **7**, 62.
57. Giabicani,E., Chantot-Bastaraud,S., Bonnard,A., Rachid,M., Whalen,S., Netchine,I. and Brioude,F. (2019) Roles of type 1 insulin-like growth factor (IGF) receptor and IGF-II in growth regulation: evidence from a patient carrying both an 11p paternal duplication and 15q deletion. *Front. Endocrinol.*, **10**, 263.
58. Chiesa,N., De Crescenzo,A., Mishra,K., Perone,L., Carella,M., Palumbo,O., Mussa,A., Sparago,A., Cerrato,F., Russo,S. *et al.* (2012) The KCNQ1OT1 imprinting control region and non-coding RNA: new properties derived from the study of Beckwith–Wiedemann syndrome and Silver–Russell syndrome cases. *Hum. Mol. Genet.*, **21**, 10–25.
59. Corpet,A., Kleijwegt,C., Roubille,S., Juillard,F., Jacquet,K., Texier,P. and Lomonte,P. (2020) PML nuclear bodies and chromatin dynamics: catch me if you can!. *Nucleic Acids Res.*, **48**, 11890–11912.
60. Shi,M., You,K., Chen,T., Hou,C., Liang,Z., Liu,M., Wang,J., Wei,T., Qin,J., Chen,Y. *et al.* (2021) Quantifying the phase separation property of chromatin-associated proteins under physiological conditions using an anti-1, 6-hexanediol index. *Genome Biol.*, **22**, 229.
61. Zhang,H., Zhao,R., Tones,J., Liu,M., Dilley,R.L., Chenoweth,D.M., Greenberg,R.A. and Lampson,M.A. (2020) Nuclear body phase separation drives telomere clustering in ALT cancer cells. *Mol. Biol. Cell*, **31**, 2048–2056.
62. Fonin,A.V., Silonov,S.A., Fefilova,A.S., Stepanenko,O.V., Gavrilova,A.A., Petukhov,A.V., Romanovich,A.E., Modina,A.L., Zueva,T.S., Nedelyaev,E.M. *et al.* (2022) New evidence of the importance of weak interactions in the formation of PML-bodies. *Int. J. Mol. Sci.*, **23**, 1613.
63. Fonin,A.V., Silonov,S.A., Shpironok,O.G., Antifeeva,I.A., Petukhov,A.V., Romanovich,A.E., Kuznetsova,I.M., Uversky,V.N. and Turoverov,K.K. (2021) The role of non-specific interactions in canonical and ALT-associated PML-bodies formation and dynamics. *Int. J. Mol. Sci.*, **22**, 5821.
64. Lu,S., Ye,Q., Singh,D., Cao,Y., Diedrich,J.K., Yates,J.R., Villa,E., Cleveland,D.W. and Corbett,K.D. (2021) The SARS-CoV-2 nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein. *Nat. Commun.*, **12**, 502.
65. Jack,A., Ferro,L.S., Trnka,M.J., Wehri,E., Nadgir,A., Nguyenla,X., Fox,D., Costa,K., Stanley,S., Schaletzky,J. *et al.* (2021) SARS-CoV-2 nucleocapsid protein forms condensates with viral genomic RNA. *PLoS Biol.*, **19**, e3001425.
66. Papa,G., Borodavka,A. and Desselberger,U. (2021) Viroplasms: assembly and functions of rotavirus replication factories. *Viruses*, **13**, 1349.
67. Levone,B.R., Lenzken,S.C., Antonaci,M., Maiser,A., Rapp,A., Conte,F., Reber,S., Mechtersheimer,J., Ronchi,A.E., Mühlemann,O. *et al.* (2021) FUS-dependent liquid–liquid phase separation is important for DNA repair initiation. *J. Cell Biol.*, **220**. e202008030.
68. Nozawa,R.-S., Yamamoto,T., Takahashi,M., Tachiwana,H., Maruyama,R., Hirota,T. and Saitoh,N. (2020) Nuclear microenvironment in cancer: control through liquid-liquid phase separation. *Cancer Sci.*, **111**, 3155–3163.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkad580/7223586 by guest on 04 August 2023

69. Trnka,F., Hoffmann,C., Wang,H., Sansevrino,R., Rankovic,B., Rost,B.R., Schmitz,D., Schmidt,H.B. and Milovanovic,D. (2021) Aberrant phase separation of FUS leads to lysosome sequestering and acidification. *Front. Cell Dev. Biol.*, **9**, 716919.
70. Levone,B.R., Lombardi,S. and Barabino,S.M. (2022) Laser microirradiation as a tool to investigate the role of liquid-liquid phase separation in DNA damage repair. *STAR Protocols*, **3**, 101146.
71. Ding,D.-Q., Okamasa,K., Katou,Y., Oya,E., Nakayama,J.-i., Chikashige,Y., Shirahige,K., Haraguchi,T. and Hiraoka,Y. (2019) Chromosome-associated RNA–protein complexes promote pairing of homologous chromosomes during meiosis in Schizosaccharomyces pombe. *Nat. Commun.*, **10**, 5598.
72. Nakagawa,S., Yamazaki,T. and Hirose,T. (2018) Molecular dissection of nuclear paraspeckles: towards understanding the emerging world of the RNP milieu. *Roy. Soc. Open Biol.*, **8**, 180150.
73. Baudement,M.-O., Cournac,A., Seveno,M., Parrinello,H., Reynes,C., Sabatier,R., Bouschet,T., Yi,Z., Sallis,S., Tancelin,M. *et al.* (2018) High-salt–recovered sequences are associated with the active chromosomal compartment and with large ribonucleoprotein complexes including nuclear bodies. *Genome Res.*, **28**, 1733–1746.
74. Grosch,M., Ittermann,S., Shaposhnikov,D. and Drukker,M. (2020) Chromatin-associated membraneless organelles in regulation of cellular differentiation. *Stem Cell Reports*, **15**, 1220–1232.
75. Bernard,D., Prasanth,K.V., Tripathi,V., Colasse,S., Nakamura,T., Xuan,Z., Zhang,M.Q., Sedel,F., Jourdren,L., Coulpier,F. *et al.* (2010) A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression. *EMBO J.*, **29**, 3082–3093.
76. Dion,W., Ballance,H., Lee,J., Pan,Y., Irfan,S., Edwards,C., Sun,M., Zhang,J., Zhang,X., Liu,S. *et al.* (2022) Four-dimensional nuclear speckle phase separation dynamics regulate proteostasis. *Sci. Adv.*, **8**, eabl4150.
77. Hughes,J.R., Roberts,N., McGowan,S., Hay,D., Giannoulatou,E., Lynch,M., De Gobbi,M., Taylor,S., Gibbons,R. and Higgs,D.R. (2014) Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.*, **46**, 205–212.
78. Beagrie,R.A., Scialdone,A., Schueler,M., Kraemer,D.C., Chotalia,M., Xie,S.Q., Barbieri,M., de Santiago,I., Lavitas,L.-M., Branco,M.R. *et al.* (2017) Complex multi-enhancer contacts captured by genome architecture mapping. *Nature*, **543**, 519–524.
79. Quinodoz,S.A., Ollikainen,N., Tabak,B., Palla,A., Schmidt,J.M., Detmar,E., Lai,M.M., Shishkin,A.A., Bhat,P., Takei,Y. *et al.* (2018) Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell*, **174**, 744–757.
80. Namekawa,S.H., Payer,B., Huynh,K.D., Jaenisch,R. and Lee,J.T. (2010) Two-step imprinted X inactivation: repeat versus genic silencing in the mouse. *Mol. Cell. Biol.*, **30**, 3187–3205.
81. Yuan,H., Yan,M., Zhang,G., Liu,W., Deng,C., Liao,G., Xu,L., Luo,T., Yan,H., Long,Z. *et al.* (2019) CancerSEA: a cancer single-cell state atlas. *Nucleic Acids Res.*, **47**, D900–D908. .
82. Consortium,I.N. (2022) 3DGenBench: a web-server to benchmark computational models for 3D Genomics. *Nucleic Acids Res.*, **50**, W4–W12.
83. Zhou,Q., Guan,P., Zhu,Z., Cheng,S., Zhou,C., Wang,H., Xu,Q., Sung,W.-k. and Li,G. (2022) ASMdb: a comprehensive database for allele-specific DNA methylation in diverse organisms. *Nucleic Acids Res.*, **50**, D60–D71. .
84. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
85. Brown,T.B., Mann,B., Ryder,N., Subbiah,M., Kaplan,J., Dhariwal,P., Neelakantan,A., Shyam,P., Sastry,G., Askell,A. *et al.* (2020) Language models are few-shot learners. *Adv. Neur. Inf. Process. Syst.*, **33**, 1877–1901.
86. Pan,S.J. and Yang,Q. (2010) A survey on transfer learning. *IEEE Trans. Knowledge Data Eng.*, **22**, 1345–1359.
87. Runge,J. (2018) Causal network reconstruction from time series: from theoretical assumptions to practical estimation. *Chaos*, **28**, 075310.
88. Runge,J., Nowack,P., Kretschmer,M., Flaxman,S. and Sejdinovic,D. (2019) Detecting and quantifying causal associations in large nonlinear time series datasets. *Sci. Adv.*, **5**, eaau4996.
89. Mayer,R., Brero,A., Von Hase,J., Schroeder,T., Cremer,T. and Dietzel,S. (2005) Common themes and cell type specific variations of higher order chromatin arrangements in the mouse. *BMC Cell Biol.*, **6**, 44.
90. Cheutin,T. and Cavalli,G. (2012) Progressive polycomb assembly on H3K27me3 compartments generates polycomb bodies with developmentally regulated motion. *PLoS Genet.*, **8**, e1002465.
91. Schueder,F., Lara-Gutiérrez,J., Beliveau,B.J., Saka,S.K., Sasaki,H.M., Woehrstein,J.B., Strauss,M.T., Grabmayr,H., Yin,P. and Jungmann,R. (2017) Multiplexed 3D super-resolution imaging of whole cells using spinning disk confocal microscopy and DNA-PAINT. *Nat. Commun.*, **8**, 2090.
92. Schlichthaerle,T., Strauss,M.T., Schueder,F., Auer,A., Nijmeijer,B., Kueblbeck,M., Jimenez Sabinina,V., Thevathasan,J.V., Ries,J., Ellenberg,J. *et al.* (2019) Direct visualization of single nuclear pore complex proteins using genetically-encoded probes for DNA-PAINT. *Angew. Chem.*, **131**, 13138–13142.
93. Di Stefano,M., Di Giovanni,F., Pozharskaia,V., Gomar-Alba,M., Baù,D., Carey,L.B., Marti-Renom,M.A. and Mendoza,M. (2020) Impact of chromosome fusions on 3D genome organization and gene expression in budding yeast. *Genetics*, **214**, 651–667.
94. Sabinina,V.J., Hossain,M.J., Hériché,J.-K., Hoess,P., Nijmeijer,B., Mosalaganti,S., Kueblbeck,M., Callegari,A., Szymborska,A., Beck,M. *et al.* (2021) Three-dimensional superresolution fluorescence microscopy maps the variable molecular architecture of the nuclear pore complex. *Mol. Biol. Cell*, **32**, 1523–1533.
95. Hu,J., Li,X., Coleman,K., Schroeder,A., Ma,N., Irwin,D.J., Lee,E.B., Shinohara,R.T. and Li,M. (2021) SpaGCN: integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods*, **18**, 1342–1351.
96. Xu,C., Jin,X., Wei,S., Wang,P., Luo,M., Xu,Z., Yang,W., Cai,Y., Xiao,L., Lin,X. *et al.* (2022) DeepST: identifying spatial domains in spatial transcriptomics by deep learning. *Nucleic Acids Res.*, **50**, e131.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkad580/7223586 by guest on 04 August 2023

© The Author(s) 2023. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com