D950–D956 *Nucleic Acids Research, 2023, Vol. 51, Database issue* *Published online 1 November 2022*
https://doi.org/10.1093/nar/gkac957

GenomicKB: a knowledge graph for the human genome

Fan Feng[1], Feitong Tang[2,†], Yijia Gao[2,†], Dongyu Zhu[3,†], Tianjun Li[2,†], Shuyuan Yang[2,†], Yuan Yao[2], Yuanhao Huang[1] and Jie Liu[1,2,*]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, MI, USA, [2]Electrical Engineering and Computer Science, University of Michigan, MI, USA and [3]School of Information, University of Michigan, MI, USA

Received August 22, 2022; Revised October 06, 2022; Editorial Decision October 07, 2022; Accepted October 27, 2022

ABSTRACT

Genomic Knowledgebase (GenomicKB) is a graph database for researchers to explore and investigate human genome, epigenome, transcriptome, and 4D nucleome with simple and efficient queries. The database uses a knowledge graph to consolidate genomic datasets and annotations from over 30 consortia and portals, including 347 million genomic entities, 1.36 billion relations, and 3.9 billion entity and relation properties. GenomicKB is equipped with a web-based query system (https://gkb.dcmb.med.umich.edu/) which allows users to query the knowledge graph with customized graph patterns and specific constraints on entities and relations. Compared with traditional tabular-structured data stored in separate data portals, GenomicKB emphasizes the relations among genomic entities, intuitively connects isolated data matrices, and supports efficient queries for scientific discoveries. GenomicKB transforms complicated analysis among multiple genomic entities and relations into coding-free queries, and facilitates data-driven genomic discoveries in the future.

INTRODUCTION

Since the completion of the Human Genome Project (1), ever-evolving biotechnologies have enabled us to characterize the human genome from different perspectives. Consequently, many landmarking consortia have made tremendous progress towards understanding the functions of human genome in different aspects, such as the Encyclopedia of DNA Elements (ENCODE) (2), Roadmap Epigenomics (3), Genotype-Tissue Expression (GTEx) (4) and 4D Nucleome (4DN) (5), among others. Although these consortia provided different insights at an unprecedented scale and depth, the separately-stored tabular data is inconvenient for genomic research and scientific discoveries. First, merging multi-modal data often requires joining multiple tables, which takes tremendous storage space and efforts. Second, it is challenging to reconcile multiple data sources for the same topic (e.g. enhancers annotated by ENCODE CCRE (2) and ENdb (6)). In addition, extracting information from these isolated data requires coding skills, making open science and reproducible research difficult.

To solve this problem, we build Genomic Knowledgebase (GenomicKB), which seamlessly integrates datasets and annotations related to the human genome into a knowledge graph. Knowledge graphs intuitively represent connected data entities, and have been applied to biological domains (7–12). Compared with traditional tabular-structured data stored at separate portals, GenomicKB emphasizes the relations between genomic entities at multiple resolutions and from multiple tissues and cell types. Entities from each consortium automatically and explicitly cross-link with one another in the knowledge graph without any operations such as table joining and sorting. In addition, our GenomicKB is rigorously built with well-defined schemata, identities, and ontologies to maintain the data structure, disambiguate genomic concepts, and support future extension. As a result, GenomicKB is not only flexible to adapt updates of nodes, relations, and entire data sources, but also connects with other knowledge graphs in related biomedical domains.

To support customized queries, GenomicKB is equipped with a user-friendly web portal (https://gkb.dcmb.med.umich.edu/). To the best of our knowledge, this is the first graph pattern query system for the human genome, in which a query does not necessarily start with a genomic region or a specific genomic entity. Instead, GenomicKB supports customized pattern queries such as 'finding two genes which are both related to signal transduction, locate on the same chromosome, and form ligand-receptor pairs'. As a result, GenomicKB transforms multi-modal data analysis into intuitive queries, and enables large-scale cross-modality pattern searching and learning in a highly-integrated knowledge graph. With this integrated data source and a robust data-sharing web portal, biomedical scientists can easily query, compare and investigate the high quality, high

*To whom correspondence should be addressed. Email: drjieliu@umich.edu
†The authors wish it to be known that, in their opinion, the second to sixth authors should be regarded as Joint Second Authors.

© The Author(s) 2022. 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

resolution, and comprehensive knowledge graph regarding chromatin organization, regulatory elements, epigenomic markers and transcriptional regulation in various human tissues/cell lines at multiple resolutions.

MATERIALS AND METHODS

Overall system design

Software systems. To build our knowledge graph system, we choose *property graphs*, one special type of data graphs, which specifies genomic entities (e.g. genes) as graph nodes, their relations (e.g. gene regulatory network) as edges, and additional information (e.g. gene descriptions) as node and edge properties. For implementation, we used *neo4j* (13), a native graph database, which efficiently implements the property graph model directly down to the storage level. The system also comes with an efficient query language (i.e. Cypher (14)) which supports constant-time traversals for both depth and breadth queries over data graphs.

Data importing. The original tabular datasets come from a number of data portals and web servers (Supplementary Note 1). We use a Python tool to automatically convert the tabular data to csv matrices which are supported by *neo4j* data loading API. For each source, a format file is used to identify data entries from the table, and help *neo4j* represent them as nodes and edges (Supplementary Note 2).

Version control and backup. Our data managing module periodically backups the backend graph database on a bi-weekly basis. Each backup is securely stored on our server and logged along with the backup date and size. Should in any case our backend graph database be modified or attacked unexpectedly, the data managing module rolls back the database to a previous backup version.

Web portal. GenomicKB is hosted on a web server at the Michigan Academic Computing Center at the University of Michigan. The backend uses python Flask as the server program and *neo4j* as the graph database. The server program connects to the database with the *neo4j* and *networkx* Python packages. The frontend uses *vis.js* to provide interactive network visualization of the graph database. Bootstrap framework is used to create a user interface. NGINX is used to perform the reverse proxy that passes internet requests to our server. User cookies are not collected. The web portal has passed a stress test to ensure reliable service under simultaneous access from multiple users.

Schema, identity and ontology in GenomicKB

Schema. Schemata prescribe high-level structures and semantics that the knowledge graph follows, which reduces data errors and allows reasoning over the data graph (15). In GenomicKB, we formally define the node schema and edge schema as follows. Nodes are labeled with hierarchical classes. The top level includes six classes, namely chromosome chain, coding element, non-coding element, epigenomic feature, variant, and ontology. Each class also consists of sub-classes (Supplementary Table S1). Edge schema defines the rules of node connections. Edges are categorized into position, regulation, expression, and annotation types, and each sub-type has corresponding start and end node types (Supplementary Table S2). For example, an 'express in' edge must start from a gene and point to a tissue or cell line, and a 'correlate with' edge only corresponds to the correlation between variants and gene expression or phenotype. Node schema and edge schema are exactly followed during data importing to ensure GenomicKB's structure, semantics and data types.

Identity. Identity consolidates a set of unique identifiers and disambiguates different genomic identities in the knowledge graph. Since different data sources may follow different conventions to represent the same concept (e.g. ENSG00000223972 and gene *DDX11L1*), or use the same name to describe different concepts (e.g., gene *p53* and protein *p53*), we use *globally-unique identifiers* and *external identity links* in GenomicKB. For example, for genes, transcripts and exons, we refer to Ensembl (16) IDs for their external identity links. For epigenomic entities without external identity links such as ChIP-seq peaks, we define their globally-unique identifiers according to their genomic coordinates, cell lines, and histone/TF types.

Ontology. Ontology is a uniform language to describe scientific terms. Concepts such as cell lines and tissues are represented as ontology URLs and IDs instead of common names to ensure disambiguity and future integration with other knowledge graphs. GenomicKB includes well-established ontologies related to genes (GO (17) and HGNC (18)), tissues and cell lines (UBERON (19), BTO (20), CL (21), and EFO (22)). These ontologies serve two roles in GenomicKB. First, some entities directly connect to ontologies and are accessible in queries. For example, users can query all genes linked to the same specified GO term. Second, scientific terms such as diseases and cell line names are encoded in ontology IDs. Therefore, different conventions of the same concept, such as 'IMR-90', 'IMR90' and 'cells-cultured fibroblasts' are unified in GenomicKB.

Graph query implementation

The workflow of a GenomicKB query includes two steps: (i) translation from a user's query graph into Cypher (14) (GenomicKB's backend query language) and (ii) returning the query result from the Cypher query. In the first step, the query graph returned from the web portal is split into '(start node) – (edge label) – (end node)' triples, and each triple is translated into Cypher. For example, '*variant* correlate_with *gene*' is translated into 'MATCH (n1:variant)-[:correlate_with]->(n2:gene) RETURN n1, n2'. Particularly, positional relations between two nodes (e.g. overlap) are converted to the positional relations between each node and the chromosome backbone (Supplementary Note 1). For example, '*gene* overlap *gene*' is translated into 'MATCH (n1:gene)-[:locate_on]->(:chr_chain)<-[:locate_on]-(n2:gene) RETURN n1, n2'. In the second step, Cypher from all triples is submitted to the neo4j query system to retrieve query results. When users submit queries on the web portal, we only return the first 5–20 matched patterns on the result page to reduce the query time (by adding

a 'LIMIT' clause to Cypher). When users click 'export all' on the result page to export complete results, GenomicKB re-submits the Cypher query but returns all matched patterns from the graph database.

RESULTS

GenomicKB integrates data from over 30 credible sources

Our knowledge graph integrates over 30 well-established data sources, including GENCODE (23), the Eukaryotic Promoter Database (EPD) (24), dbSuper (26), RNAcentral (25), Genotype-Tissue Expression (GTEx) (4), GWAS (27), Database of Genomic Variants (DGV) (28), NCBI dbVar (29), 4D Nucleome (4DN) (5), FIRE studies (30), ENCODE (2), MotifMap (31), NCBO ontologies (32), etc. (Supplementary Tables S1 and S2). Each of these consortia incorporates thousands of datasets and provides different insights regarding human genome at an unprecedented scale and depth. Information is explicitly represented as nodes, edges and properties in GenomicKB, resulting in 347 378 103 nodes, 1 359 209 258 edges and 3 902 460 300 node/edge properties. To the best of our knowledge, the coverage of GenomicKB exceeds any knowledge graphs in related fields (7–12). One vital advantage of our knowledge graph structure is its flexibility which allows easy inclusion of new data in different formats. In addition, the query efficiency only drops insignificantly as we increase data entries (Supplementary Note 3).

GenomicKB supports graph-based queries

We design a web portal (http://gkb.dcmb.med.umich.edu/) that supports customized queries of diverse entities, relations and properties. The query system consists of a canvas, an editor panel, and a console. On the canvas, users can draw customized graph patterns by adding nodes and edges. When adding a node/edge or a node/edge is selected, the corresponding editor panel on the top left activates to enable node/edge configuration, such as editing the type of the node/edge or adding property constraints. During the process, the console shows real-time hints to guide users to create valid queries. After the user specifies the query conditions, the user needs to click the 'Submit' button on the bottom to submit the query, which re-directs to a result page (Supplementary Note 5).

The result page includes two panels. The left panel displays the result sub-graph with moving and zooming functions. If positional relations (such as *overlap* and *downstream*) are included in the query, genomic regions that entities locate in are also visualized as connected bins, whereas other entities related to this region are displayed around it. The right panel displays detailed properties when a node is selected. If the retrieved sub-graph is overly large, then only partial results (e.g. five to twenty matched patterns) are visualized, and the complete query result can be downloaded by clicking 'export all'. The downloaded result is in excel format. A video tutorial is also available on our front page.

GenomicKB simplifies cross-modality analysis as queries over the knowledge graph

GenomicKB integrates complementary data sources into a knowledge graph and simplifies multi-modal analysis as queries over the knowledge graph. For example, to identify genes and genetic variants related to type II diabetes (T2D), traditional approaches require integrating multiple data sources as follows. First, all variants correlated with T2D are retrieved from portals such as GWAS Catalog (27). Then, variants are linked to genes by identifying intra-gene variants with gene coordinates from GENCODE (23). Additional restrictions about the gene may be applied as well, such as the minimum gene expression level in pancreas (from consortia such as ENCODE (2) and GTEx (4)). Lastly, function annotations of the genes are identified from Gene Ontology (17). With GenomicKB, the aforementioned analysis can be easily completed with a sub-graph query over the knowledge graph (Figure 1). All restrictions and sub-graphs can be specified via the user-friendly interface, and the system no longer require complex queries in individual data sources or any coding skills. At the backend, the submitted query pattern is automatically translated into a Cypher query (14), and the query results are returned and visualized as graphs (Figure 1). With consolidated data and an intuitive query process, GenomicKB makes it easier for researchers to discover new genomic insights.

GenomicKB encodes positional relations among different genomic entities

Most genomic entities locate on specific regions on the chromosome with positional relations between each other. GenomicKB supports queries based on positional relations including *locate_in* (one entity is completely included by another), *overlap* (two entities have a coordinate overlap), *upstream*/*downstream* (one entity does not overlap and is upstream/downstream of another on the same chromosome), and *same_chr* (two entities are on the same chromosome). For example, to investigate transcription factor (TF) binding at chromatin loop anchors called at 5 kb resolution, the traditional approach is to call loops from chromatin contact maps available at 4DN data portal and collect TF binding profile from epigenome consortia such as ENCODE and Roadmap Epigenomics, and then identify their overlap with computational tools. In GenomicKB, a query '*TF_binding_site* overlap *loop*' provides the same result (Figure 2). When restricting the query to GM12878 cell line and the TF name to be CTCF, 4724 distinct loops are returned. As a comparison, 5758 are returned from the query of all GM12878 loops without specifying the overlap with TF binding sites. Therefore, 82% of loops in GM12878 have at least one anchor bound by CTCF. A similar query of loops overlapping two different CTCF binding sites results in 2,680 returned entries, indicating that 47% of the 5758 loops are between two CTCF binding sites.

To represent positional relations in GenomicKB, we first split all chromosomes into regions of a particular size (i.e. resolution), represent each region as a node, and connect them with edges. The series of nodes and edges are referred to as 'chromosome chains', which are constructed

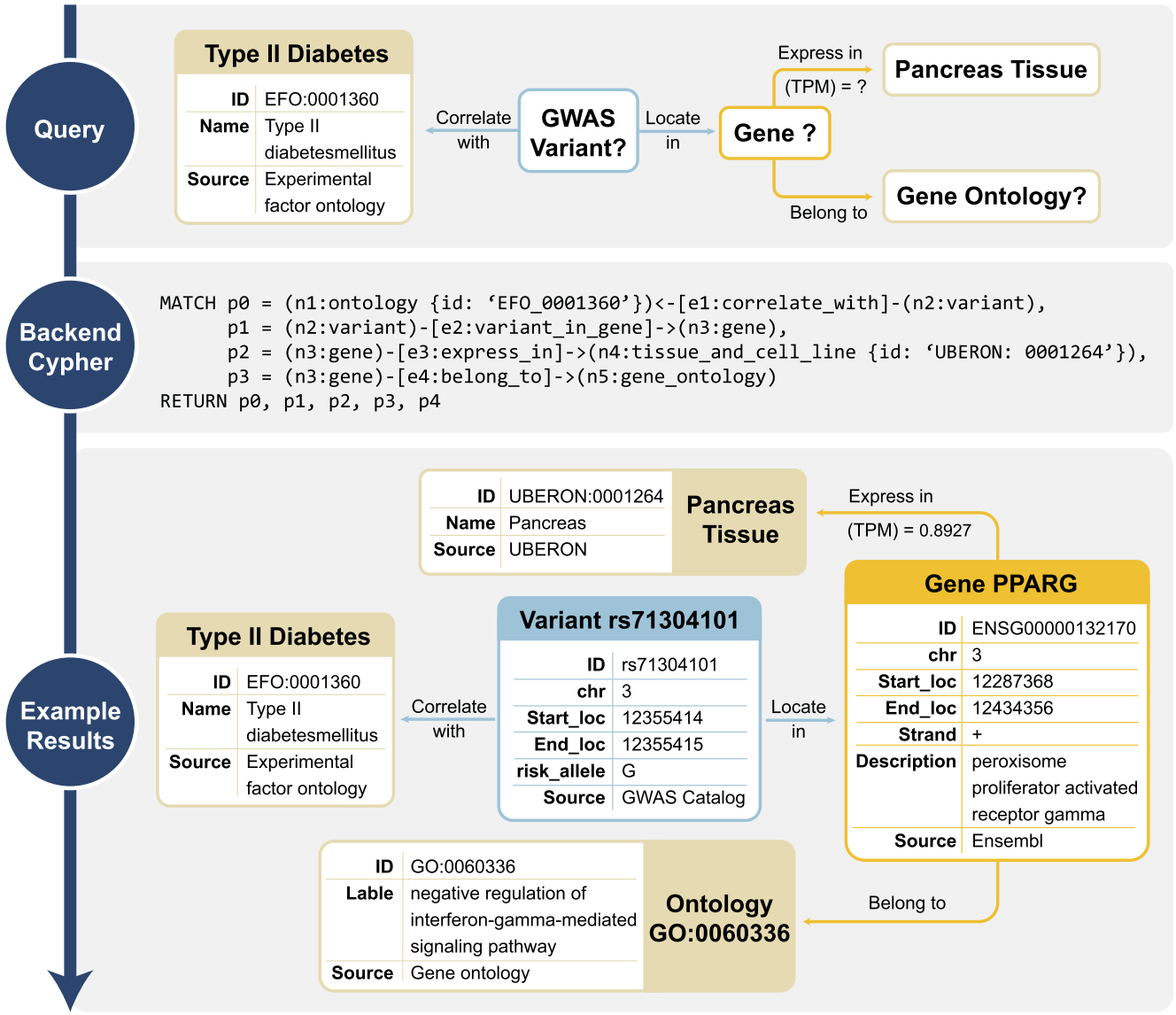

Figure 1. GenomicKB simplifies cross-modality analysis as queries over the knowledge graph. If a user is interested in relations between T2D and genes, then instead of searching multiple databases including GWAS, ENCODE and GO, a sub-graph query over GenomicKB returns all variants, genes and gene ontologies that satisfy the query criteria.

in different resolutions. Afterwards, entities that locate on specific regions are connected to the corresponding chromosome chain nodes. The chromosome chains are intermediate nodes for capturing any positional relations between genomic entities (Figure 2, details in Supplementary Note 1).

GenomicKB reconciles consensus or conflicting data sources of the same problem

For some genomic entity, multiple data sources may provide either consensus or conflicting evidence. Knowledge graphs are able to reconcile duplicate or conflicting facts in the light of well-defined schemata, identities, and ontologies. We use the example of enhancers to show that GenomicKB reconciles multiple data sources for the same domain. As key regulatory elements, enhancers are annotated by several data sources, such as ENdb (6), EnhancerAtlas (33), ENCODE CCRE (2) and FANTOM5 (34). To identify enhancers from one database in GenomicKB, users can query the node 'enhancer' with restrictions such as 'data_source = FANTOM5'. By defining enhancers from different data sources with coordinate overlaps as consensus ones, one can also query how many enhancers from two sources (e.g., CCRE and EnhancerAtlas) agree with each other (Query 1 in Figure 3). In addition, relations from one data source can be cross-validated by other data sources. For example, EnhancerAtlas provides enhancer-gene interactions, which can be validated by other approaches that map enhancers to genes such as eQTL-gene correlation as follows. First, a query '*enhancer* regulate *gene*' with restriction 'cell_line=GM12878' and

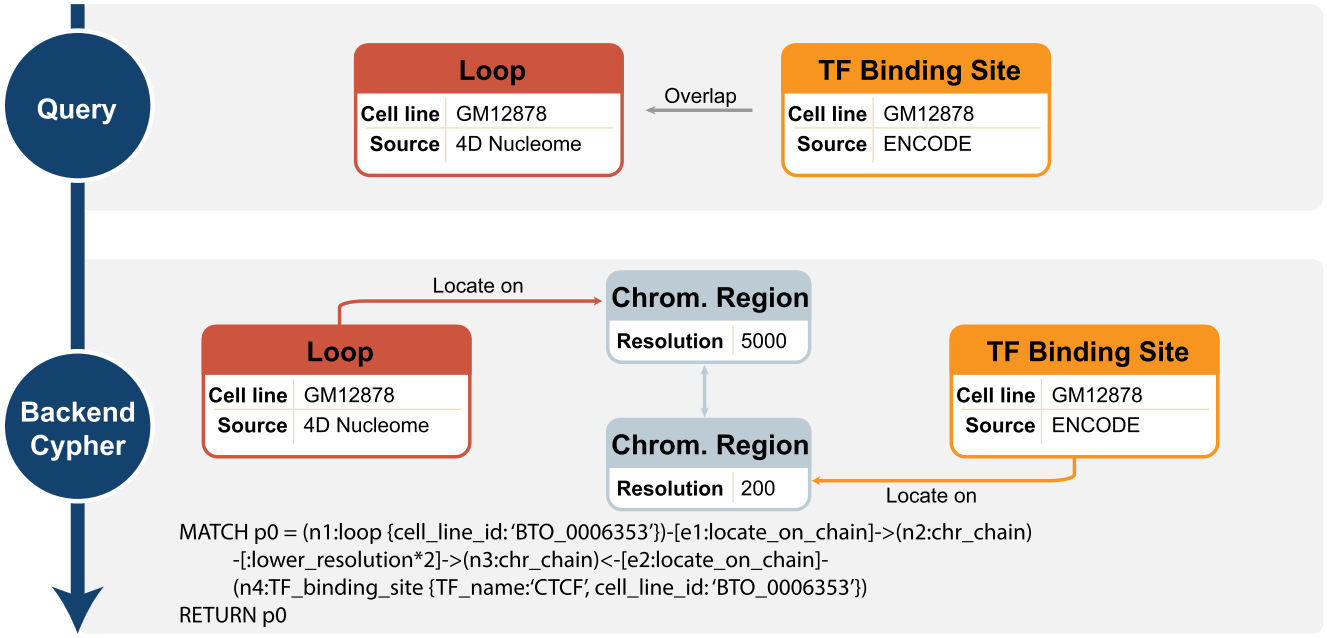

Figure 2. GenomicKB supports queries related to positional relations between genomic entities. An example query of CTCF binding to loop anchors is illustrated.

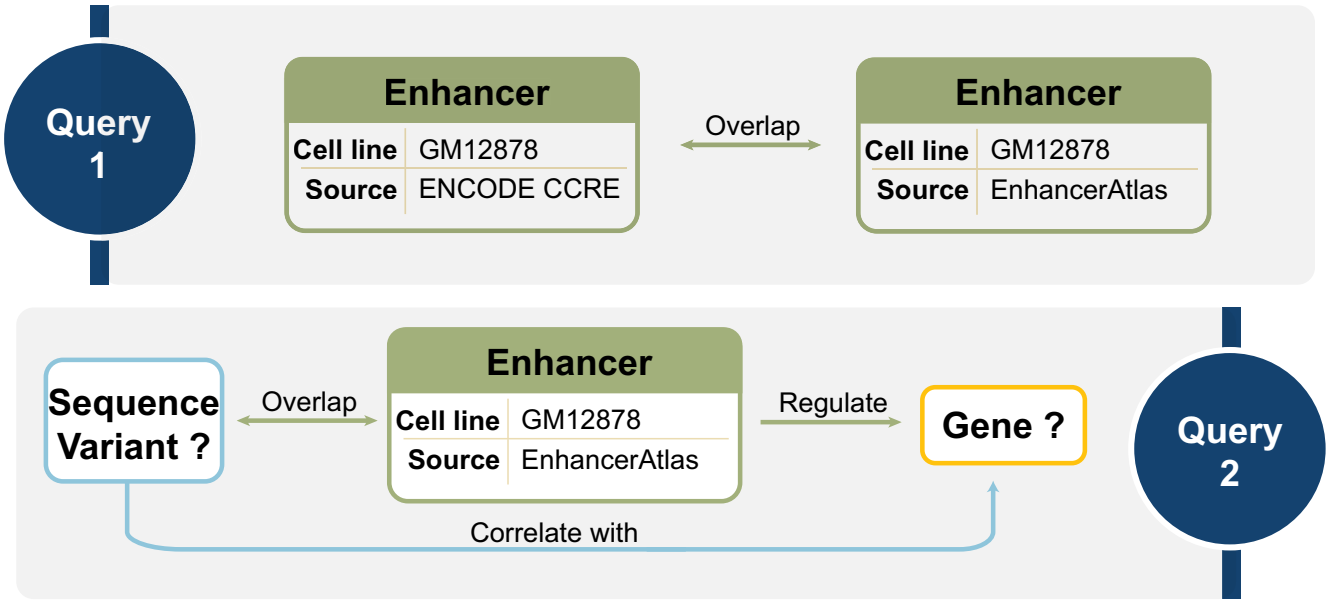

Figure 3. GenomicKB reconciles multiple data sources for the same problem, such as identifying enhancers and mapping enhancers to genes. Query 1 demonstrates how GenomicKB evaluates the consensus enhancers between CCRE and EnhancerAtlas. Query 2 illustrates how enhancer-gene mapping from EnhancerAtlas is validated by eQTL-gene pairs in GenomicKB.

'data_source=EnhancerAtlas' returns 118,610 enhancer-gene pairs from EnhancerAtlas. Then, we can identify the eQTLs of the gene locating in the enhancer, which can be represented as '*variant* overlap *enhancer*', '*enhancer* regulate *gene*', and '*variant* correlate_with *gene*' (Query 2 in Figure 3). The number of distinct enhancer-gene pairs decreases to 16 871 in the result, indicating that 16 871 enhancer-gene pairs from EnhancerAtlas can be validated by GTEx eQTLs.

DISCUSSION

In conclusion, GenomicKB integrates our existing knowledge regarding human genome, epigenome, transcriptome, and 4D nucleome in a large knowledge graph. Different from traditional tabular-structured data, it emphasizes the relations between different perspectives and provides explicit connections between entities of interest. With the flexibility, well-defined schemata and ontologies used in the knowledge graph, it is quite easy to update the existing

entities and relations and incrementally add more entities and relations. Since GenomicKB adapts external unique identifiers for nodes and edges, it is convenient to connect it with other biomedical knowledge graphs. To increase accessibility, GenomicKB is equipped with a web portal (http://gkb.dcmb.med.umich.edu/) for users to specify and submit intuitive graph-based queries. With this portal, GenomicKB is capable of answering human genomics-related questions and conducting multi-modal analysis with a coding-free and interactive queries. Therefore, we expect that GenomicKB can attract researchers with diverse backgrounds and promote open science in genomic research.

In recent years, artificial intelligence plays increasingly important roles in problems related to transcription regulation (35–38), chromatin 3D structures (39–42), and single-cell genomics (43,44). Nevertheless, we are still looking for a 'universal model' that captures large-scale genomic data from different perspectives and comprehensively decodes the human genome. Similar to the field of natural language processing in which new language models and question-answering systems are based on large knowledge graphs (45,46) (e.g. the Wiki knowledge graph), we expect that genomic research becomes increasingly data-driven, and GenomicKB provides high-quality and integrated data for large-scale machine learning methods and facilitates scientific discoveries.

DATA AVAILABILITY

All data used in this study are publicly available. Source data are provided with this paper (Supplementary Note 1, Supplementary Tables 1 and 2).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

NIH [R35HG011279 to J.L.]. Funding for open access charge: NIH [R35HG011279 to J.L.].
Conflict of interest statement. None declared.

REFERENCES

1. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **40**, 860–921.
2. Moore,J.E., Purcaro,M.J., Pratt,H.E., Epstein,C.B., Shoresh,N., Adrian,J., Kawli,T., Davis,C.A., Dobin,A., Kaul,R. *et al.* (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, **583**, 699–710.
3. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F., Ren,B., Milosavljevic,A., Meissner,A., Kellis,M., Marra,M.A., Beaudet,A.L., Ecker,J.R. *et al.* (2010) The NIH roadmap epigenomics mapping consortium. *Nat. Biotech.*, **28**, 1045.
4. GTEx Consortium (2017) Genetic effects on gene expression across human tissues. *Nature*, **55**, 204–213.
5. Dekker,J., Belmont,A.S., Guttman,M., Leshyk,V.O., Lis,J.T., Lomvardas,S., Mirny,L.A., O'shea,C.C., Park,P.J., Ren,B. *et al.* (2017) The 4D nucleome project. *Nature*, **549**, 219.
6. Bai,X., Shi,S., Ai,B., Jiang,Y., Liu,Y., Han,X., Xu,M., Pan,Q., Wang,F., Wang,Q. *et al.* (2020) ENdb: a manually curated database of experimentally supported enhancers for human and mouse. *Nucleic Acids Res.*, **48**, D51–D57.
7. Santos,A., Colaço,A.R., Nielsen,A.B., Niu,L., Strauss,M., Geyer,P.E., Coscia,F., Albrechtsen,N.J.W., Mundt,F., Jensen,L.J. *et al.* (2022) A knowledge graph to interpret clinical proteomics data. *Nat. Biotech.*, **40**, 1–11.
8. Yoon,B.H., Kim,S.K. and Kim,S.Y. (2017) Use of graph database for the integration of heterogeneous biological data. *Genom. Inform.*, **15**, 19.
9. Balaur,I., Mazein,A., Saqi,M., Lysenko,A., Rawlings,C.J., Auffray,C. *et al.* (2017) Recon2Neo4j: applying graph database technologies for managing comprehensive genome-scale networks. *Bioinformatics*, **33**, 1096–1098.
10. Himmelstein,D.S., Lizee,A., Hessler,C., Brueggeman,L., Chen,S.L., Hadley,D., Green,A., Khankhanian,P. and Baranzini,S.E. (2017) Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife*, **6**, e26726.
11. Mughal,S., Moghul,I., Yu,J., Clark,T., Gregory,D.S. and Pontikos,N. (2017) Pheno4J: a gene to phenotype graph database. *Bioinformatics*, **33**, 3317–3319.
12. Barabási,A.L., Gulbahce,N. and Loscalzo,J. (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.*, **12**, 56–68.
13. Webber,J. (2012) A programmatic introduction to neo4j. In: *Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity*. pp. 217–218.
14. Francis,N., Green,A., Guagliardo,P., Libkin,L., Lindaaker,T., Marsault,V., Plantikow,S., Rydberg,M., Selmer,P. and Taylor,A. (2018) Cypher: an evolving query language for property graphs. *Proceedings of the 2018 International Conference on Management of Data*. 1433–1445.
15. Hogan,A., Blomqvist,E., Cochez,M., D'Amato,C., Melo,G.D., Gutierrez,C., Kirrane,S., Gayo,J.E.L., Navigli,R., Neumaier,S. *et al.* (2021) Knowledge graphs. *ACM Comput. Surv. (CSUR)*, **54**, 1–37.
16. Howe,K.L., Achuthan,P., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R., Bhai,J. *et al.* (2021) Ensembl 2021. *Nucleic Acids Res.*, **49**, D884–D891.
17. Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
18. Povey,S., Lovering,R., Bruford,E., Wright,M., Lush,M. and Wain,H. (2001) The HUGO gene nomenclature committee (HGNC). *Human Genet.*, **109**, 678–680.
19. Mungall,C.J., Torniai,C., Gkoutos,G.V., Lewis,S.E. and Haendel,M.A. (2012) Uberon, an integrative multi-species anatomy ontology. *Genome Biol.*, **13**, 1–20.
20. Gremse,M., Chang,A., Schomburg,I., Grote,A., Scheer,M., Ebeling,C. and Schomburg,D. (2010) The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res.*, **39**, D507–D513.
21. Diehl,A.D., Meehan,T.F., Bradford,Y.M., Brush,M.H., Dahdul,W.M., Dougall,D.S., He,Y., Osumi-Sutherland,D., Ruttenberg,A., Sarntivijai,S. *et al.* (2016) The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J. Biom. Semant.*, **7**, 1–10.
22. Malone,J., Holloway,E., Adamusiak,T., Kapushesky,M., Zheng,J., Kolesnikov,N., Zhukova,A., Brazma,A. and Parkinson,H. (2010) Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, **26**, 1112–1118.
23. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A. and Searle,S. (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *gr*, **22**, 1760–1774.
24. Dreos,R., Ambrosini,G., Cavin Périer,R. and Bucher,P. (2013) EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. *Nucleic Acids Res.*, **41**, D157–D164.
25. RNAcentral Consortium (2021) RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res.*, **49**, D212–D220.
26. Khan,A. and Zhang,X. (2015) dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res.*, **44**, D164–D171.
27. MacArthur,J., Bowler,E., Cerezo,M., Gil,L., Hall,P., Hastings,E., Junkins,H., McMahon,A., Milano,A., Morales,J. *et al.* (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.*, **45**, D896–D901.

28. MacDonald,J.R., Ziman,R., Yuen,R.K., Feuk,L. and Scherer,S.W. (2014) The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.*, **42**, D986–D992.
29. Lappalainen,I., Lopez,J., Skipper,L., Hefferon,T., Spalding,J.D., Garner,J., Chen,C., Maguire,M., Corbett,M., Zhou,G. *et al.* (2012) DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res.*, **41**, D936–D941.
30. Schmitt,A.D., Hu,M., Jung,I., Xu,Z., Qiu,Y., Tan,C.L., Li,Y., Lin,S., Lin,Y., Barr,C.L. and Ren,B. (2016) A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Rep.*, **17**, 2042–205.
31. Daily,K., Patel,V.R., Rigor,P., Xie,X. and Baldi,P. (2011) MotifMap: integrative genome-wide maps of regulatory motif sites for model species. *BMC Bioinformatics*, **12**, 1–13.
32. Martínez-Romero,M., Jonquet,C., O'connor,M.J., Graybeal,J., Pazos,A. and Musen,M.A. (2017) NCBO Ontology Recommender 2.0: an enhanced approach for biomedical ontology recommendation. *J. Biom. Semant.*, **8**, 1–22.
33. Gao,T. and Qian,J. (2020) EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res.*, **48**, D58–D64.
34. Andersson,R., Gebhard,C., Miguel-Escalada,I., Hoof,I., Bornholdt,J., Boyd,M., Chen,Y., Zhao,X., Schmidl,C., Suzuki,T. *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature*, **507**, 455–461.
35. Zhou,J., Theesfeld,C.L., Yao,K., Chen,K.M., Wong,A.K. and Troyanskaya,O.G. (2018) Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.*, **50**, 1171–1179.
36. Kelley,D.R., Reshef,Y.A., Bileschi,M., Belanger,D., McLean,C.Y. and Snoek,J. (2018) Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.*, **28**, 739–750.
37. Kelley,D.R. (2020) Cross-species regulatory sequence activity prediction. *PLoS Comput. Biol.*, **16**, e1008050.
38. Avsec,Ž., Agarwal,V., Visentin,D., Ledsam,J.R., Grabska-Barwinska,A., Taylor,K.R., Assael,Y., Jumper,J., Kohli,P. *et al.* (2021) Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods*, **18**, 1196–1203.
39. Fudenberg,G., Kelley,D.R. and Pollard,K.S. (2020) Predicting 3D genome folding from DNA sequence with Akita. *Nat. Methods*, **17**, 1111–1117.
40. Belokopytova,P.S., Nuriddinov,M.A., Mozheiko,E.A., Fishman,D. and Fishman,V. (2020) Quantitative prediction of enhancer–promoter interactions. *Genome Res.*, **30**, 72–84.
41. Zhang,S., Chasman,D., Knaack,S. and Roy,S. (2019) In silico prediction of high-resolution Hi-C interaction matrices. *Nat. Commun.*, **10**, 1–18.
42. Li,W., Wong,W.H. and Jiang,R. (2019) DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic Acids Res.*, **47**, e60–e60.
43. Fu,L., Zhang,L., Dollinger,E., Peng,Q., Nie,Q. and Xie,X. (2020) Predicting transcription factor binding in single cells through deep learning. *Sci. Adv.*, **6**, eaba9031.
44. Ma,Q. and Xu,D. (2022) Deep learning shapes single-cell data analysis. *Nat. Rev. Mol. Cell Biol.*, 1–2.
45. Chen,X., Jia,S. and Xiang,Y. (2020) A review: knowledge reasoning over knowledge graph. *Expert Syst. Appl.*, **141**, 112948.
46. Nicholson,D.N. and Greene,C.S. (2020) Constructing knowledge graphs and their biomedical applications. *Comput. Struct. Biotech. J.*, **18**, 1414–1428.