Bioinformatics, 2026, **42**(3), btag013
https://doi.org/10.1093/bioinformatics/btag013
Published: 20 January 2026
Original Paper | Data and text mining

OXFORD

Aggregation of gene regulatory information and knowledge on FAIR principles enables discovery of pathogenic gene regulatory variants

Keyang Yu[1,†], Haoquan Zhao[1,†], Andrea S. Wilderman[1], Tierra R. Farris[1], Jessie E. Arce[1], David Chen[1], Andrew R. Jackson[1], Yiran Guo[2], Qi Li[2], Bosko Jevtic[3], Dubravka Jevtic[3], Vuk Milinovic[3], Yuankun Zhu[2], Jeremy Costanza[4], Eric D. Wenger[2], Christopher Nemarich[2], Lisa Anderson[5], Aleksandar Mihajlović[3], Kristin Ardlie[5], Shaine A. Morris[6], Matthew E. Roth[1], Deanne M. Taylor[2,7], Adam C. Resnick[2,7], Lilei Zhang[1], Aleksandar Milosavljevic[1,*]

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, 77030, United States
[2]Center for Data Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, United States
[3]Persida Inc., Brooklyn, NY, 11219, United States
[4]Azrieli Research Center, CHU Sainte-Justine Research Centre, Montreal, QC, H3T 1C5, Canada
[5]Broad Institute of MIT and Harvard, Cambridge, MA, 02142, United States
[6]Department of Pediatrics, Baylor College of Medicine and Texas Children's Hospital, Houston, TX, 77030, United States
[7]Department of Biomedical and Health Informatics; Department of Pediatrics, The Children's Hospital of Philadelphia, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, 19104, United States
*Corresponding author. Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, 400D, Houston, TX, 77030, United States. E-mail: amilosav@bcm.edu
[†]= equal contribution.
Associate Editor: Jonathan Wren

Abstract

Motivation: **Methods for sharing gene regulatory information and knowledge on FAIR principles, particularly in the context of tissue-specific gene regulation, remain poorly defined and implemented, hampering discovery and clinical genetic diagnosis.**

Results: **We specified FAIR principles for tissue-specific gene regulatory information and knowledge; implemented them by developing a registry of regulatory elements and aggregating FAIR gene regulatory information from several major sources; developed computational tools that utilize these FAIR resources; and demonstrated their utility by associating gene regulatory variants with major subtypes of congenital heart disease.**

Availability and implementation: **Variant prioritization infrastructure tools are available in genboree node repository at https://genboree.org/verdaccio/#/. Detailed documentation is available at https://ldh.clinicalgenome.org/docs/ldh/overview.html#related-services. The code for use case analyses and free access variant data is available on Zenodo with DOI: https://doi.org/10.5281/zenodo.17833070.**

1 Introduction

Dramatic reduction of whole genome sequencing (WGS) cost (Mardis 2011) and the growing coverage of WGS by health insurance plans (Trosman *et al.* 2020, Phillips *et al.* 2022) have accelerated adoption of WGS for clinical diagnosis and research purposes (Borchert *et al.* 2021, Hodder *et al.* 2024). The transition from diagnostic whole exome sequencing (WES) to WGS brings into view the genetic variants outside of the ∼1% of the human genome that codes for amino acids. Such noncoding variants may be functional and pathogenic because they regulate gene expression (Blake *et al.* 2023).

WGS is often applied if WES does not yield genetic diagnosis. A good example is congenital heart diseases (CHD), a frequent condition affecting ∼1% of live births. WES and chromosomal microarray identify pathogenic variants for less than 30% of the cases (Zaidi and Brueckner 2017). Recently, however, the increased accessibility of WGS has enabled researchers to identify

Received: 3 July 2025. **Revised:** 10 December 2025. **Accepted:** 5 January 2026

regulatory variants that contribute to the condition (Smemo *et al.* 2012). Seminal study in 2020 suggests that about one in every 1000 newborns may carry a regulatory variant that causes CHD (Richter *et al.* 2020). Moreover, CHD is often caused by dominant *de novo* variants (DNVs) that can be detected by trio WGS and more readily identified as pathogenic due to their relatively small number per individual.

The identification of pathogenic regulatory variants is challenging because of the large number of gene regulatory elements (enhancers, insulators, promoters, etc.), exceeding by orders of magnitude the number of protein coding genes (Coppola *et al.* 2016). The ENCODE (Dunham *et al.* 2012, Luo *et al.* 2020) project alone produced a map of over a million candidate gene regulatory elements over 20% of the genome (Lee and Young 2013, Levine *et al.* 2014, Oudelaar and Higgs 2021, Preissl *et al.* 2023). Unlike the effects of protein coding variants within the context of the classical triplet code and changes in protein structure, the effects of gene regulatory variants are highly dependent on the specific regulatory element affected, specific cell and tissue type, and developmental context (Moore *et al.* 2020, The GTEx Consortium 2020).

The interpretation of noncoding variants is informed by many layers of information about gene regulation. A foundational map of gene regulatory elements in developmental context across 111 cell and tissue types was provided by the NIH Roadmap Epigenome Project (Kundaje *et al.* 2015, Onuchic *et al.* 2018). GTEx published a rich source of information about tissue-specific gene expression and eQTL associations (The GTEx Consortium 2020). The ongoing developmental GTEx (dGTEx) project expands the early maps to include relations of variants, regulatory elements, and genes in developmental, cellular, and anatomical contexts (Coorens *et al.* 2025). Other related projects such as the Human Cell Atlas (Rozenblatt-Rosen *et al.* 2017), PsychEncode (Emani *et al.* 2024), and the Brain Initiative Cell Atlas Network (BICAN) are applying an increasing diversity of assays to assess gene expression and decipher gene regulation within specific cell types and tissues in various developmental and physiological contexts.

The increasing diversity of sources and volume of data about gene regulation calls for aggregation strategies that are more scalable than the classical centralized data warehousing method (Grigoriu *et al.* 2021, Orlov *et al.* 2023). When it comes to raw data and metadata, FAIR (Findable, Accessible, Interoperable, and Reusable) principles (Wilkinson *et al.* 2016) have seen significant adoption (Garcia *et al.* 2019, Kapoor *et al.* 2024) and improvements (Wilkinson *et al.* 2018), and have proven effective. This is amplified by the advent of AI tools, which critically depend on AI-readiness of training data, FAIRness being a key requirement for AI-readiness (Alharbi *et al.* 2024, Clark *et al.* 2024).

While FAIRness has been traditionally defined in the context of data, the interpretation of WGS does not rely just on the raw data but also on the information and knowledge about developmental gene regulation. For example, the most used product of GTEx is not the raw data produced by the project but the derived information about tissue-specific gene expression and quantitative trait loci (xQTL) information linking gene regulatory variants to gene expression. This calls for extending FAIR principles beyond just the original data, which was the traditional focus of FAIRness.

To extend the scope of application of FAIR principles, working within the framework of Ackoff's Data, Information, and Knowledge pyramid (Ackoff 1989), we define *Information* as being derived from data by analysis methods and *Knowledge* as consisting of general assertions supported by one or more lines of evidence. Examples of Knowledge are assertions about pathogenicity of variants, tissue-specific expression of genes, and activity of specific regulatory elements in specific cellular and anatomical contexts. Within this framework, we extend FAIR principles beyond just the data and associated metadata to also include Information and Knowledge about gene regulation and present four distinct contributions. First, working within the NIH Common Fund Data Ecosystem (CFDE), we developed a consensus RFC document by employing the CFDE RFC process (Supplementary file 1, available as supplementary data at *Bioinformatics* online) (Charbonneau *et al.* 2022). Second, we implemented the RFC by developing a registry for regulatory elements and by linking gene regulatory information across major public sources via the CFDE Linked Data Hub. Third, we provided exemplary user interfaces for regulatory variant discovery and visualization that utilize the aggregated information. Finally, we validated the utility of the aggregated information by identifying *de novo* gene regulatory variants that are associated with major subtypes of congenital heart disease.

2 Materials and methods

2.1 CFDE Request for Comment (RFC) development

Following the Request for Comment (RFC) process defined by CFDE, we drafted the RFC document that addressed FAIR aggregation of gene regulatory data, information, and knowledge. Briefly, a questionnaire was developed to gather input from CF data coordination centers (DCCs) on the type of Data, Information, and Knowledge hosted, shared, and used as well as the FAIR principles adopted (Table 1, available as supplementary data at *Bioinformatics* online). The results from 4D Nucleome, exRNA, GlyGen, GTEx, HuBMAP, IDG, Kids First, LINCS, Metabolomics, and MoTrPAC DCCs were collected, and a "Team-draft" of the RFC was developed based on the results and sent to the CFDE-RFC-Admin for review. The approved document was converted into a "CFDE-Draft" for distribution to the whole consortium to gather input and feedback. A final publication of the RFC document, defined as the "CFDE-RFC," was approved by the CFDE Admin and made available to the community as a "Standard Implementation" document. All CFDE RFCs are available at: https://github.com/nih-cfde/rfcs/blob/master/adoptionstatus.md.

2.2 reGL Registry service

The regulatory element Genomic Location (reGL) Registry service aims to provide unique identifiers (reGLids) for genomic regions with transcriptional or post-transcriptional regulatory roles. The reGL service is supported by two more generic

services, the Genomic Location (GL) Registry and the Alias service of the Linked Data Hub. The former was implemented in node.js with ArangoDB as a backend database and provides reference-agnostic GLids for genomic regions. The LDH-Alias service provides reGLids for regulatory elements and maps them to GLids and other commonly used identifiers through an LDH-like subject-to-linked data model. Node.JS packages for both services are available under MIT license at our node repository (https://genboree.org/verdaccio/#/). Requests for bulk or individual registration of new regulatory element loci should be sent to brl-glreg-help@bcm.edu.

2.3 Regulatory elements

The largest number of cis-regulatory elements (cREs) that were assigned reGL identifies using the reGL registration service ($n = 1\,063\,878$) originate from the SCREEN v3 database. They are supported by ENCODE annotations of DNase activity, H3K4me3 modification, H3K27ac modification, and CTCF-binding activity. Regions with promoter-like signatures (PLS) and enhancer-like signatures (ELS), as defined by SCREEN, were selected for analysis within this study. Disease-specific regulatory elements were selected based on the phenotype-associated tissues in which the cRE was called. Additional regulatory elements, including developmental cis-regulatory elements from publications, and eCLIP-identified RNA-binding protein targeted sites, were assigned reGL identifies using the reGL registration service. Only regulatory elements with clear regulatory state (active/inactive) or defined regulatory type were included in this study.

2.4 Common Fund Data Ecosystem Gene Regulation Linked Data Hub (CFDE LDH)

The CFDE LDH service was implemented using ArangoDB. Customized RESTful APIs were designed based on ArangoQL queries to retrieve documents from the database in bulk.

2.5 Kids First Data Resource Portal and Variant WorkBench

Kids First Data Resource Center (https://kidsfirstdrc.org/) developed the Kids First Data Resource Portal (KFDRP: https://portal.kidsfirstdrc.org/), which is a centralized data platform for both Kids First and collaborative cohorts. To assist disease-related variants identification in the human genome, KFDRP released the Spark-based Variant WorkBench (VWB), based on Cavatica (https://cavatica.sbgenomics.com/). VWB hosts harmonized genomic variational and curated phenotypic/clinical information for released Kids First cohorts.

2.6 Variant calling

2.6.1 Patients

WGS and phenotypic data for congenital heart diseases trios ($n = 763$) were obtained from the Pediatric Cardiac Genetics Consortium (PCGC) Congenital Heart Disease Network Study (phs001138.v3.p2, phs001194.v3.p2). The DNVs for the 763 trios were called using previously described pipeline (Richter et al., 2020). Supplementary variant data for 517 out of the 763 probands were obtained from the Kids First Data Resource Portal (KFDRP, phs001194.v4.p3). Variants were called by using the Kids First DRC Joint Genotyping Workflow (https://github.com/kids-first/kf-jointgenotyping-workflow). Data annotation and DNV filtering were performed on CAVATICA using the KFVWB. All variants were registered through the ClinGen Allele Registry (Pawliczek *et al.* 2018) service. Probands were mapped to the PCGC phenotype data (phs001194.v3.p2) and categorized into five CHD subtypes based on clinical standards: left ventricular obstruction track defect, conotruncal defect, laterality defect; right-sided defects, and atrial septal defect.

2.6.2 Controls

DNVs from unaffected siblings were called based on 1611 sibling–parent trios from sporadic autism quartets that consist of one affected sibling, one unaffected sibling, and unaffected parents. Variant calling for control WGS data was performed together with the CHD cohort as previously described (Richter et al., 2020).

2.7 CHD genes

We constructed a focus gene set with 22 genes that have previously been identified to harbor pathogenic coding variants in hypoplastic left heart syndrome (HLHS), one of the most common LVOT defects. This gene set included NOTCH1, which has been found to harbor several low-penetrance variants alongside FOXC2, FOXL1, NKX2-5, and HAND1 and their candidate regulatory regions (Iascone *et al.* 2012). Additional candidates were drawn from families with HLHS, and reported cardiomyopathy. MYBPC3, RYR2, and MYH6 were identified with inherited variants, and interaction between MYH6 and FLNC was also described (Theis *et al.* 2021). MYO15A was included based on findings from a separate cohort with HLHS (Theis and Olson 2022). Rare variants in NEXN, CASQ2, SOS1, TTN, TXNRD2, DSG2, FOXRED1, SCN5A, LAMA4, SYNE1, and GJA5 were reported in HLHS cases by Helle *et al.* (2020). Some of these genes may relate to additive or combinatorial effects of inherited and de novo variation, as supported by animal models (Yagi *et al.* 2018, Liu *et al.* 2019). A more comprehensive list of 1114 CHD genes for subtype-specific gene prioritization analysis were obtained from https://chddb.fwgenetics.org/ (Zhou *et al.* 2023).

2.8 Distance model-based burden test

Distribution of the distances from variants to their closest cRE centers were compared between patients and healthy controls for a given gene set. All DNVs and cREs within the gene body or the 10 kbps upstream and 10 kbps downstream region of the genes were included in the test. Rank sum tests were used to measure the overall difference in distribution, and the Kolmogorov–Smirnov (KS) tests were used to measure the pointwise largest difference in distribution. For gene-based burden tests, Rank sum tests and KS tests were performed for each gene individually. For subtype-specific models, only cREs active in subtype-specific tissues and cell types were included in the model.

3 Result

3.1 RFC document to guide FAIR sharing of regulatory data, information, and knowledge

To develop the RFCs document (Supplementary file 2, available as supplementary data at *Bioinformatics* online), we followed the CFDE RFC framework and process. We first assessed the scope of gene regulatory Data, Information, and Knowledge within the CFDE consortium. This included evaluating the current implementation of the FAIR (Findable, Accessible, Interoperable, and Reusable) principles. The following 10 data coordination centers (DCCs) within CFDE were surveyed: GTEx, 4D Nucleome, exRNA, GlyGen, HuBMAP, IDG, Kids First, LINCS, Metabolomics, and MoTrPAC. All but one DCC reported sharing gene-associated data, whereas only four shared variant-associated data. HuBMAP, MoTrPAC, and 4D Nucleome shared data associated with regulatory regions, though none linked it directly to both genes and variants. Assessment of FAIR compliance revealed that gene and variant identifiers varied across DCCs, and no standardized identifier was used for regulatory elements. Only six of the 10 DCCs provided Information or Knowledge about genes or variants, with limited inclusion of disease-associated Data. These observations indicate incomplete implementation of FAIR principles particularly in relation to Information and Knowledge about regulatory elements, which we addressed in the RFC.

We next reviewed relevant standards. We found that the Global Alliance for Genomics and Health (GA4GH) FAIR standards, such as the Variant Representation Specification (VRS) (Goar *et al.* 2023), refget APIs (Yates *et al.* 2021), Workflow Execution Service (WES), and Data Connect (Thorogood *et al.* 2021) were widely implemented within the field. Other efforts, including the Ensembl API (Yates *et al.* 2020) and SmartAPI (Zaveri *et al.* 2017), addressed similar goals. Nevertheless, most of these standards do not address the specifics of gene regulatory Information and Knowledge. To that end, we created an RFC that concentrated on three core entities—genes, variants, and regulatory elements—and addressed each FAIR principle in relation to Information and Knowledge associated with these entities and their relations.

To support Findability, the RFC specifies naming conventions and attributes required for each entity. Identifiers must link to de-referencing services that return key attributes based on a defined genome assembly. For regulatory elements, the RFC recommends the use of ENCODE-SCREEN identifiers (Moore *et al.* 2020) as a *de facto* community standard. We address by developing a regulatory element registry, as described below. For Accessibility, the RFC recommends REpresentational State Transfer Application Programming Interfaces (RESTful APIs) to enable retrieval of Data, Information, and Knowledge via the HTTP protocol. Compared to alternatives such as GraphQL or WebSocket, RESTful HTTP APIs are more readily implemented, and their use in conjunction with JSON provides flexible integration using modern web programming standards. The RFC explicitly recommends JSON or JSON-LD for API output to support both human and machine readability.

For Reusability, the RFC defines two record types: “Evidence” (representing Information) and “Assertion” (representing Knowledge), where each Assertion must be supported by one or more Evidence entries. Both must reference Findable identifiers corresponding to relevant entities. To connect the assertions to lines of evidence, the RFC recommends the use of the Scientific Evidence and Provenance Information Ontology (SEPIO) (Brush *et al.* 2016), which models relationships between scientific claims, supporting evidence, and underlying data. The RFC also recommends the use of ontology terms to standardize the vocabulary used across records, particularly for communicating the context (cell and tissue type), which are very important for associating noncoding variants with specific subtypes of developmental disorders such as congenital heart disease (Fig. 1A). Specifically, the RFC recommends using Uberon (Mungall *et al.* 2012) for tissue/organ annotations and the Cell Ontology (CL) (Diehl *et al.* 2016) for cell types. Other ontologies can be incorporated, but mapping to Uberon and CL are preferred per the RFC. Cross-ontology harmonization could be performed using a workflow combining lexical matching, OWL 2 DL (the Web Ontology Language in its Description Logic profile) reasoning, and expert manual curation (Mungall *et al.* 2012) to reconcile terms and preserve provenance.

3.2 Implementing FAIR principles in the domain of gene regulation by leveraging ClinGen infrastructure

In prior work with the ClinGen project, we addressed FAIR implementation for genomic variant Knowledge by developing the ClinGen Allele Registry (CAR) (Pawliczek *et al.* 2018). CAR is a web-based service that generates globally unique, canonical variant identifiers (CAids) via RESTful APIs. The identifiers are reference agnostic and represent the same variant from across GRCh38, GRCh37, and NCBI 36 human reference genomes. The system performs de-duplication dynamically, independent of database submission workflows, and returns either HTML or JSON-LD documents to external tools (Table 1). CAR supports the registration and standardized naming of a range of variant types, including single nucleotide polymorphisms (SNPs), small insertions and deletions (indels), and structural variants (SVs). Built in alignment with GA4GH Variant Representation Specifications (VRS), CAR integrates with reference variant databases such as ClinVar (Landrum *et al.* 2025), myVariant.info (Xin *et al.* 2016), and CIViC (Griffith *et al.* 2017). In addition, the service maintains a versioned database of gene identifiers, cross-referenced with HGNC (Seal *et al.* 2023), Entrez (NCBI Resource Coordinators 2013), and Ensembl gene symbol repositories (Yates *et al.* 2020).

While variant registration has seen increasing alignment with FAIR principles, the problem of naming new regulatory elements that are not in the SCREEN database persists. To address this gap, we developed a system for assigning identifiers to regulatory elements (reGLids.). This system is implemented as an aliasing service based on the generic Genomic Location Registry (GLR) that assigns identifiers (GLids) to any genomic locus. One-to-one aliasing is implemented using the Linked Data Hub Alias service (LDH-Alias). GLR, built on a multi-model ArangoDB framework, supports efficient indexing and retrieval of positional data associated with genomic intervals (Table 1).

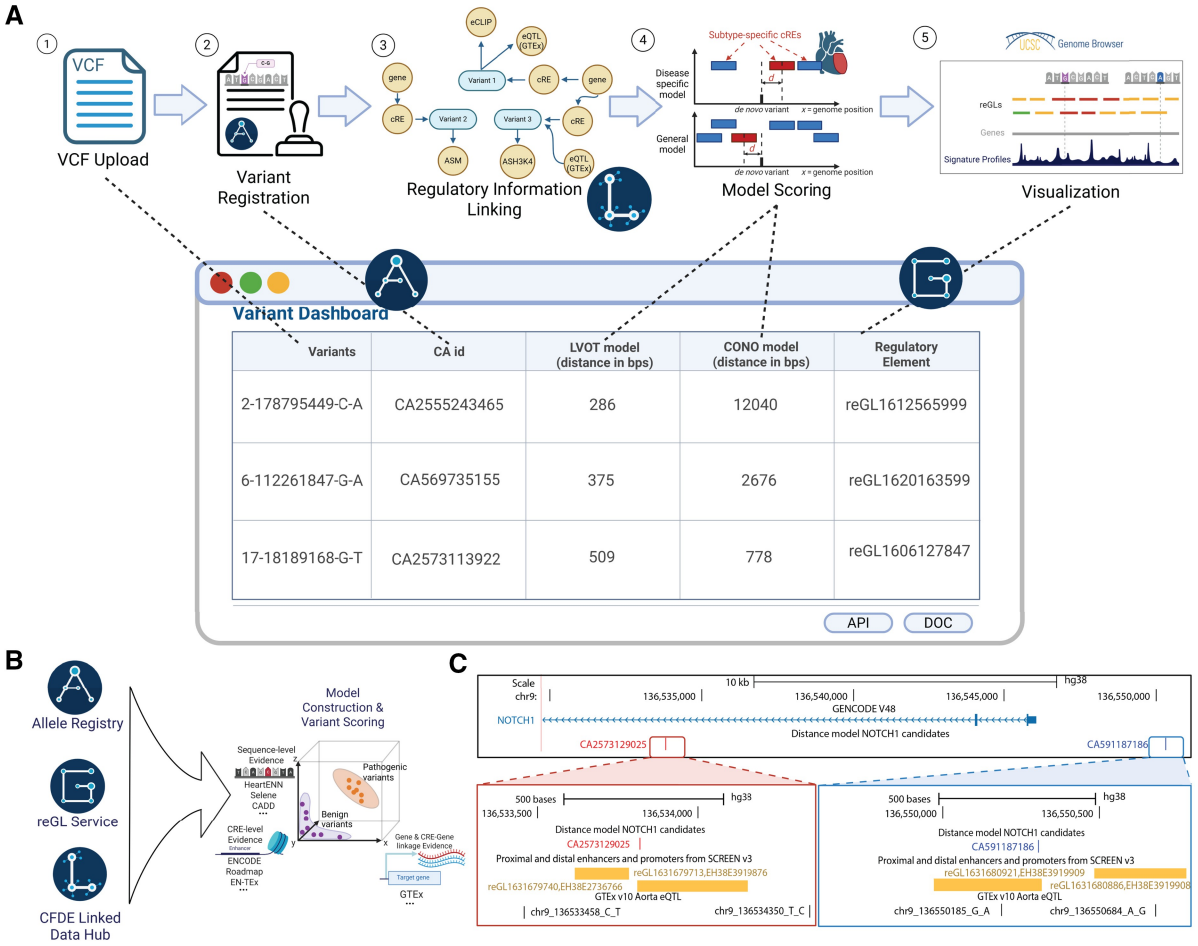

Figure 1 FAIR infrastructure for regulatory variant interpretation and prioritization. (A) The FAIR framework supports interpretation and prioritization of candidate noncoding variants. (B) The ClinGen Allele Registry, Genomic Location Registry, and CFDE Linked Data Hub link enables structured integration of regulatory data, information, and knowledge through standardized identifiers and APIs. (C) Visualization of the dataset facilitates assessment of reGLs based on external databases. Panel A and B created with BioRender.com.

Table 1 Implementation of FAIR principles for gene regulatory information and knowledge.

Layer	Findable	Accessible	Interoperable	Reusable
Knowledge	Unique and persistent ids Indexed Linked to contributing information	RESTful HTTP APIs Follow OpenAPI specifications Registered at SmartAPI	JSON formatted documents	Provenance in SEPIO-like data structure
Information	Unique and persistent ids Indexed Linked to variants, genes, REs			UBERON, Cell, and Cell Line ontology terms URL to information source

Genomic regions can be registered and named by any registered entity, when a chromosome, start, and end position are defined. The system returns stable, reference-agnostic Genomic Location identifiers (GLids) on request and maintains their mapping across relevant reference assemblies and supports positional queries, thus eliminating the need for "annotation lifting" and enabling integration of information and knowledge across sources that utilize different reference assemblies (Table 1).

"reGLids" serve as context-specific aliases, stored in LDH-Alias as key-value pairs. Each reGLid maps 1:1 to a GLid and can be further connected to additional aliases (such as identifiers in SCREEN and other databases) when relevant. The service

Table 2 UI and API documentations for tools within the infrastructure.

FAIR infrastructure		UI and API documentations
Clingen Allele Registry	**UI**	https://reg. genome.network.org/
	API	https://reg. genome.network.org/docs/cg-car/endpoints
CFDE Linked Data Hub	**UI**	https://ldh.genome.network/cfde/ldh/ui/
	API	https://ldh.genome.network/cfde/ldh/srvc (https://ldh.genome.network/cfde/ldh/ui/api-docs)
reGL Registry service	**API**	https://ldh.genome.network/ldh/alias/genereg/srvc (https://genboree.org/docs/ldh-alias/)
Tools		**UI and API documentations**
Variant Dashboard	**UI**	https://genboree.org/gvi/vd-ui/app
	API	https://genboree.org/docs/gvi/dst
UCSC Track Hub	**UI**	https://genboree.org/pub/genereg/ucsc_trackhub/hub.txt (for use in https://genome.ucsc.edu/cgi-bin/hgHubConnect)

supports bidirectional mapping between reGLids and associated identifiers, enabling consistent resolution and referencing across systems, databases, and reference assemblies.

All identifiers generated by CAR, GLR, and LDH-Alias, including variant and region-level aliases and genomic coordinates, are accessible through RESTful HTTP APIs (Tables 1 and 2). Functional extensions support batch operations, cross-referencing, and detection of overlapping regions. To support Interoperability, all API responses are returned in JSON format, facilitating integration into computational workflows and analysis pipelines (Table 1). FAIRness for the aforementioned tools was regularly evaluated using FAIRshake as well as other FAIRness assessment tools (Clarke *et al.* 2019, Bonello *et al.* 2022).

3.3 CFDE linked data hub connects regulatory data, information, and knowledge across data silos

With standardized identifiers in place for gene-associated features involved in regulation, we then implemented a framework for the exchange of gene regulatory Information and Knowledge. While regulatory relationships span a broad range of contexts, our initial use case focused on the interpretation of regulatory noncoding variants and identification of associated disease-relevant elements, with a focus on DNVs by applying a distance-based collapsing burden test to identify pathogenic variants that are overrepresented in affected individuals relative to unaffected controls. While previously applied to coding regions, we adapted burden testing to associate noncoding DNVs that occur in closer proximity to regulatory elements relevant to the subtype of disease observed in individual probands. The approach leverages the knowledge about genes previously associated with congenital heart disease (CHD) and information about the activity of their regulatory elements in the cell and tissue types that are relevant for subtypes of CHD.

To aggregate regulatory information on FAIR principles specified in the RFC, we created the CFDE Linked Data Hub (LDH) instance. Like the original ClinGen LDH instance (developed for aggregating information and knowledge about genetic variants for ClinGen purposes), the CFDE LDH is based on ArangoDB, a multi-model graph database capable of storing entities and their relationships as subjects (genes, regulatory elements, variants) and collections of "linked data" about them (Table 1).

To connect regulatory information across sources, we first registered relevant genomic features, including regulatory elements from SCREEN, genes from GTEx, and variants from EN-TEx, using the CAR and GLR services. These reference-agnostic identifiers that map genomic features across references enable consistent integration across datasets despite differences in coordinate systems. In the CFDE LDH registry, regulatory elements are linked to all genes within 10 kilobases of the element. Variants located in a regulatory element therefore inherit all candidate gene links via that element. We do not collapse these to a single "best" gene in the database. Context-dependent regulatory roles for regulatory elements and variants were captured by linking each feature to one or more Evidence documents that describe supporting observations. These documents include statistical summaries such as effect size, epigenetic signal levels, and significance values, along with the corresponding tissue or cell type. In GTEx, where expression and splicing QTLs (eQTLs and sQTLs) were associated with specific genes, evidence was linked to multiple features to reflect regulatory interactions.

Queries initiated on any genomic feature return associated evidence from multiple sources, generating connected knowledge graphs to support hypothesis development. Both user interface and RESTful API endpoints support retrieval in machine-readable JSON format (Table 1). The system also supports formation of higher order Knowledge records, built by linking multiple lines of evidence that support the same assertion of regulatory relationship. This structure enables ongoing knowledge refinement as new studies contribute additional evidence to the database. All documents include source references to support provenance tracking.

The CFDE LDH currently integrates Information and Knowledge from over 10 sources, including CAR, the Kids First Data Resource Center (KFDRC), SCREEN v3 (Dunham *et al.* 2012, Luo *et al.* 2020), HGNC (Seal *et al.* 2023), Roadmap Epigenome (Kundaje *et al.* 2015, Onuchic *et al.* 2018), EN-TEx (Rozowsky *et al.* 2023), GTEx (The GTEx Consortium 2020), ENCODE (Luo *et al.* 2020), and other published datasets (VanOudenhove *et al.* 2020) (Table 3). The database

Table 3 CFDE Linked Data Hub content and sources.

Category	Entity type	Source	Current entity count
Core entities	Gene	HGNC	18 701
	Regulatory elements	SCREEN, eCLIP, Publications	2 M
	Variants	CAR, Kids First	242 M
Regulatory evidence	ENCODE regulatory element evidence	SCREEN, ENCODE	654.6 M
	xQTL evidence	GTEx, Publications	8.4 M
	Allele specific evidence	EN-TEx, Roadmap	10 M
	RBP regulatory evidence	ENCORE eCLIP	937 K

supports large-scale integration, with close to one billion genomic features and evidence excerpts, submitted via a RESTful API-based messaging system. These records span more than 1000 tissue and cell types, offering a comprehensive platform for investigating regulatory variation in disease-relevant cellular contexts.

3.4 FAIR regulatory information and knowledge facilitate the discovery of disease-causing variants and regulatory elements

With the integration of FAIR-aligned services including GLR, LDH-Alias, and cross-database regulatory information via the CFDE LDH, we validated those resources by showing how they can support identification of pathogenic regulatory elements and noncoding variants in disease-focused studies. In the following, we apply the infrastructure in two distinct use cases.

3.4.1 Variant scoring based on aggregated information reveals variants and regulatory elements associated with CHD

In this use case, we explored how regulatory information retrieved from CFDE LDH can be used to identify top scoring variants that fall within the range of known disease-associated genes. Using WGS data from 763 individuals with CHD (Table 2, available as supplementary data at *Bioinformatics* online), provided by PCGC, we identified 58 203 DNVs. Each variant was registered in CAR and mapped to the nearest gene and regulatory element based on genomic coordinates. Prior studies suggested that the sensitivity to genetic perturbation diminishes with the distance from centers of annotated enhancers and transcription start sites (Onuchic *et al.* 2018). We therefore hypothesized that proximity to an active cRE, including enhancers and promoters, would predict pathogenicity and incorporated this distance metric into a burden testing framework. Rather than assessing burden on a per-gene basis, we tested the overall enrichment of regulatory DNVs in the CHD cohort that fall within short distances of the cREs that are likely relevant for specific subtype of CHD based on aggregated evidence.

We focused the analysis on 22 autosomal genes previously implicated in CHD (Table 3, available as supplementary data at *Bioinformatics* online), particularly left ventricular outflow tract (LVOT) defects (Table 4, available as supplementary data at *Bioinformatics* online), and regulatory elements within 10 kilobases of the genes, filtering based on the activity of regulatory elements within etiologically relevant cell types. We constructed a CHD-specific burden model focusing on cREs active in cardiomyocytes, one of the cell types relevant for CHD. Promoter-like sequences (PLS) and enhancer-like sequences (ELS) were identified based on regulatory signatures in the LDH excerpts. Tissue- and cell type-specific evidence supporting regulatory activity was retrieved from the LDH through API queries targeting the SCREEN collection. Additional regulatory features, such as eQTLs from GTEx overlapping these cREs, were obtained using cRE- and variant-specific API endpoints.

The filtering produced a total of 1388 candidate cREs in proximity to the 22 CHD genes. Among these, 319 were active in at least one left ventricular or cardiac muscle cell sample, based on LDH queries. Using DNVs from the PCGC CHD cohort, we tested for enrichment of DNVs near active cREs. Compared to the control cohort, DNVs in CHD cases showed marginally closer proximity to cardiomyocyte-active cRE centers (rank sum test, $P < .1$) (Fig. 2A). We further examined the distance of 1 Kbps within which we see the strongest shift and observed an enriched set of eight candidate pathogenic DNVs (enrichment ratio = 3.7; Fisher's exact test, $P = .02$; Table 5, available as supplementary data at *Bioinformatics* online). This enrichment and proximity shift were not detected when non–tissue-specific cREs were used (rank sum test, $P = .76$; enrichment ratio = 0.98; Fisher's exact test, $P = 1$) (Fig. 2B). These results suggest that LDH-derived regulatory information enables prioritization of likely pathogenic DNVs in disease cohorts, reducing the need for genome-wide scans and allowing disease- and tissue-specific analysis of the noncoding genome.

3.4.2. Tissue-specific regulatory information associates genes with specific subtypes of CHD

Like several other developmental disorders, CHD is heterogeneous, with specific lesions being distinguishable prenatally by *in utero* echocardiography (Sun 2021). The conditions may be caused by mutations in different sets of genes, which are in turn regulated by regulatory elements with varied activity levels in specific cell-type and developmental contexts (Ma *et al.* 2024). Because specific smaller gene sets and regulatory contexts correlate to a significant, yet to be fully understood, degree with specific subtypes of CHD (Morton *et al.* 2022), the advances in

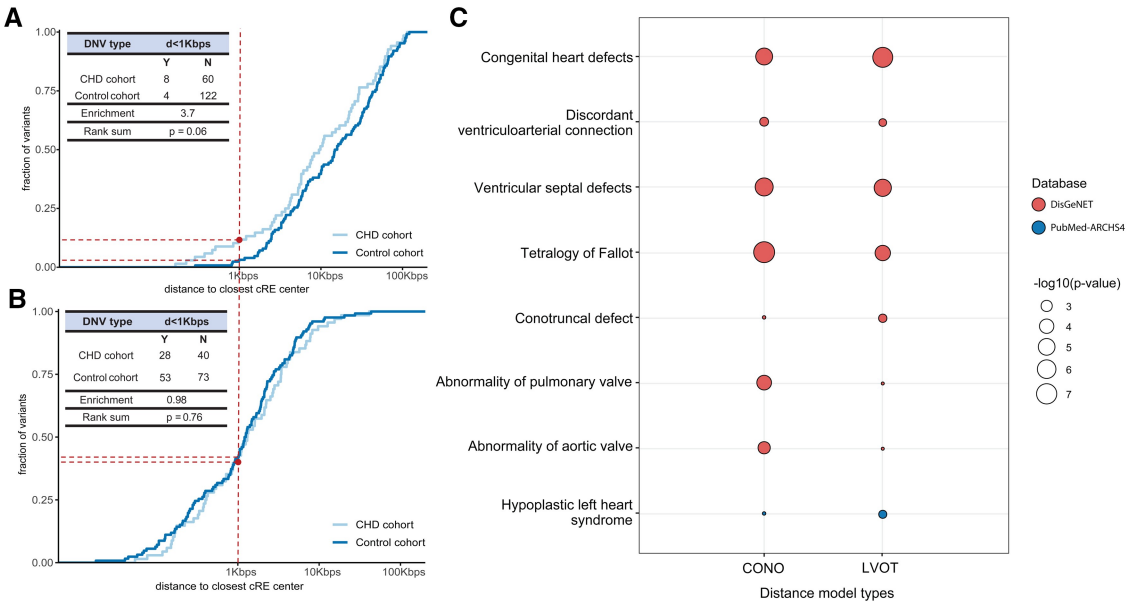

Figure 2 Burden tests revealed candidate disease causative DNVs and subtype-specific gene enrichment in CHD. (A) Burden test showed significant excess of DNVs in proximity to potentially disease-relevant cREs when comparing DNVs from a CHD cohort and a control cohort, each cohort containing over 500 WGS profiles (Wilcoxon rank sum, $P = .06$). Excess of DNVs within 1 Kbp range was highly significant (Fisher's exact test, $P = .02$). (B) Burden test showed no excess of DNVs when tissue-specific cREs were not selected. (C) Gene set enrichment analysis shows subtype-specific disease associations. Top ranking genes identified by CONO distance model are enriched for "Tetralogy of Fallot" (adj. $P = 9.33e^{-5}$) using DisGeNET annotations, while top ranking genes identified by the LVOT distance model show enrichment for "Hypoplastic left heart syndrome" (adj. $P = .037$) based on PubMed and ARCHS4 data.

CHD phenotyping are important in identifying pathogenic gene regulatory variants.

As the activity of the regulatory elements changes during development and depends on specific cellular and anatomical context, we hypothesized that pathogenic variants in specific regulatory elements should be associated with specific subtypes of CHD. Under that assumption, we asked, in this use case, if the information about cell-type specific activity of regulatory elements may help identify the genes implicated in specific subtypes of CHD. To answer this question, we constructed targeted burden models for two clinically defined subtypes of CHD, left ventricular outflow tract obstruction/defect (LVOT) and conotruncal defect (CONO). For LVOT, where over half of the individuals from our cohort were diagnosed with HLHS, we employed a cardiomyocyte-focused cRE model. For CONO, which involves a broader range of developmental mechanisms, we used a composite model incorporating cREs active in multiple tissues, including arteries and right ventricle.

To identify genes with subtype-specific regulatory association, we began with a comprehensive CHD gene list from CHD Base, a curated resource derived from 1114 disease-related publications (Zhou *et al.* 2023) (Table 6, available as supplementary data at *Bioinformatics* online). For each gene, along with a ± 10 kilobase window, we performed a collapsing burden test and computed distances between observed DNVs and nearby cREs from each subtype model (Tables 4 and 7, available as supplementary data at *Bioinformatics* online). To improve the power of the test, we supplemented the DNVs from the PCGC cohort with new DNVs within the cohort identified by the Kids First project. Stringent filters were applied to remove duplicate calls across studies and low confidence DNVs with high occurrence across probands. From this analysis, we selected genes for each subtype-specific distance model that showed the most significant difference in DNV-to-cRE distance distributions between CHD cases and controls (K–S test, $P < .05$; Tables 8 and 9, available as supplementary data at *Bioinformatics* online).

Gene enrichment analysis for disease ontology showed that "Tetralogy of Fallot," a representative malformation within the conotruncal subtype, showed highly significant association (adj. $P = 9.33e{-}5$) with the top scoring genes in the CONO model based on DisGeNET data (Table 10, available as supplementary data at *Bioinformatics* online). In comparison, its association with the LVOT top genes is significantly lower (adj. $P = 2.75e{-}3$; Table 11, available as supplementary data at *Bioinformatics* online). Meanwhile, we also observed strong enrichment for "Hypoplastic left heart syndrome" (adj. $P = .037$) association with the LVOT top genes based on Pubmed search augmented with ARCHS4 RNA-seq data (Table 12, available as supplementary data at *Bioinformatics* online). The same enrichment was not present with the CONO genes (adj. $P = .25$; Table 13, available as supplementary data at *Bioinformatics* online) (Fig. 2C). These findings indicated that regulatory information can also be used to distinguish and prioritize disease subtype-specific gene sets. Moreover, the resulting gene sets can, in turn, enhance the sensitivity of regulatory noncoding variant discovery when integrated into targeted burden models.

3.5 User interfaces for regulatory variant visualization and discovery

The APIs empower the community to build user interfaces that cater to specific data analysis needs. In the following, we present a tool for visualization of gene regulatory information using UCSC track hub and a Variant Dashboard for identification of pathogenic variants in CHD and other diseases from lists of DNVs found in patients.

3.5.1 Visualization of FAIR gene regulatory information

To facilitate visualization of gene regulatory elements impacted by potentially pathogenic variants, all currently registered regulatory genomic locations (reGLs) have been consolidated into a UCSC Genome Browser track hub (https://genboree.org/pub/genereg/ucsc_trackhub/hub.txt) (Perez *et al.* 2025). The hub enables users to visualize reGLs and assess overlaps with other genome annotations. Overlaps between registered "reGLs" can be queried via a customized GLR API.

We used this feature to analyze cREs associated with the eight prioritized CHD DNVs identified in the distance-based burden model. The highest-scoring variant, CA2573129025 (chr9:g.136533819C>T), lies in the second intron of NOTCH1 and resides within the regulatory element reGL1631679740. This cRE overlaps with enhancer elements reGL764497552 and reGL748081696, previously annotated in an embryonic heart development study, supporting a regulatory role in CHD. It also coincides with post-transcriptional regulatory elements reGL561726451 and reGL561821389, identified in ENCODE eCLIP data as targets of the RNA-binding protein EWSR1 (Dunham *et al.* 2012, Luo *et al.* 2020). A nearby intronic variant, CA12976373 (chr9:g.136534457C>T), was also reported as an aortic eQTL in GTEx (The GTEx Consortium 2020), further linking this region to NOTCH1 expression regulation (Fig. 1C).

3.5.2 Variant Dashboard supports a simple variant interpretation workflow, from variant scoring to visualization

To demonstrate the utility of the FAIR information, we developed a light-weight Variant Dashboard that supports a variant analysis workflow. The Dashboard allows users to upload and register DNVs found in individual subjects and score them against available CHD subtype-specific models, overlay information about the variants from the LDH, and access visualizations of regulatory elements that are closest to most significant variants using the UCSC Track Hub.

The scoring is performed using multiple subtype-specific models, allowing the ranking of the variant-model pairs by significance. The scoring utility is made available through customized API for improved Accessibility (Table 1). Additional variant information from the LDH, such as allele frequency, will also be retrieved through API queries to further improve the pathogenicity evaluation of significant variants. The Dashboard also provides links to the Track Hub visualization tool to evaluate the most proximal regulatory element that may mediate pathogenic effect of the variant.

4 Discussion

We anticipate that the RFC document presented here will provide guidance for aggregating gene regulatory information by the research community. While the continued advances in genomic assays and data collection technologies will require evolution of specific data formats, the three core entities—variants, regulatory elements, and genes—will likely persist. Moreover, we anticipate that the distinction between data, information, and knowledge will continue to be relevant into the future, particularly with the increasing adoption of agentic AI workflows (Zhou *et al.* 2025) in research and clinical diagnosis.

We put the RFC into practice by reusing existing resources and developing new ones as needed. In contrast to genes and variants, for which canonical naming services (HGNC and CAR respectively) existed, canonical naming services for gene regulatory elements were lacking. Addressing this gap, we developed GLR service for lookup and registration of canonical reGL identifiers for gene regulatory elements. Unlike the CAR, which served as a model and source of some components, GLR does not yet provide a service for completely independent registration of new reGLs by the users. We anticipate introducing this feature in the near future, as it will become necessary for mapping locations of regulatory elements in cell-type specific and developmental stage-specific manner at ever higher resolution and throughput using single-cell methods. Currently, registration requests for new regulatory elements or other genomic regions can be submitted through email (brl-glreg-help@bcm.edu) and will be manually reviewed. Regulatory elements with validated provenance will be accepted. Response to the requests will be provided within a week. We expect this need to register genomic regions will be amplified by the newer technologies that provide increasing levels of resolution for mapping chromatin accessibility and footprints of individual transcription factors in cell-, tissue-, and development stage-specific contexts.

We demonstrated the power of aggregated gene regulatory information to reveal genetic variants in disease subtypes by creating variant scoring models. We focused on the *cis*-effect variants and regulatory regions located within gene bodies or upstream and downstream of the gene. This allowed us to examine the transcription factor binding site-rich promoter and promoter proximal regions, as well as regions close to the 5′ and 3′ UTR, which have been shown to harbor regulatory motifs (Moore *et al.* 2020). Nevertheless, the more distal regions may still participate in gene transcription regulation through chromatin looping and will warrant future investigation using spatial regulatory information.

In addition to CHD, we recognize that other developmental disorders such as specific subtypes of epilepsy, autism, attention-deficit/hyperactivity disorder, and neurocognitive disorders are thought to also be attributable in part to mutations in many genes that are developmentally regulated by thousands of regulatory elements. As the information about gene regulation accumulates, we anticipate that the approaches presented will be able to aggregate them at scale, thus enabling discoveries of disease-associated regulatory regions in developmental disorders that affect all human organs, including the brain.

To address disorders beyond the CHD, the CFDE LDH includes regulatory information context beyond heart. Different CFDE LDH instances can also be created to tackle the volume and diversity of information. By virtue of FAIRness, related information, about

the same variant or regulatory element, e.g. can be seamlessly aggregated from instance-specific API endpoints. The CAR, GLR, and LDH-Alias databases also support the registration and naming of other types of variants, including structural variants that are also shown to be etiologically important.

Leveraging access to FAIR regulatory information via APIs, we develop a tool for visualization of regulatory information using the UCSC Track Hub and a Variant Dashboard for variant interpretation. These two tools illustrate the value of FAIR information as a substrate for the development of user interfaces that cater to specific use cases. By outsourcing complex data handling via APIs to FAIR sources of information, tools such as specialized Variant Dashboards may in the future be developed automatically even by non-programmers as lightweight UI components using AI programming tools.

We anticipate that FAIRness and programmatic (API) accessibility of information and knowledge will become increasingly relevant for integration of AI into discovery workflows. Recent publications by the Bridge To AI consortium and international initiatives recognized data FAIRness as a key element of data AI-readiness (Alharbi *et al.* 2024, Clark *et al.* 2024). We envision a future where FAIRness of both the input and output of AI tools creates an ecosystem that can facilitate coordination of hybrid workflows with combined participation of humans and agentic AI, and we hope that the work presented here will help catalyze the emergence and growth of such future discovery ecosystems.

Acknowledgements

We thank the members of the Common Fund Data Ecosystem and the Clinical Genome Resource for their continued support and collaboration. WGS and variant data were generated by the Pediatric Cardiac Genomics Consortium (PCGC), under the auspices of the National Heart, Lung, and Blood Institute's Bench to Bassinet Program https://benchtobassinet.com. The PCGC program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through grants UM1HL128711, UM1HL098162, UM1HL098147, UM1HL098123, UM1HL128761, and U01HL131003. This manuscript was not prepared in collaboration with investigators of the PCGC, has not been reviewed and/or approved by the PCGC, and does not necessarily reflect the opinions of the PCGC investigators or the NHLBI. We are grateful to the Bruce Gelb laboratory for assistance with DNV calling. The results of use case 2 analyzed and published here are based in part upon data generated by Gabriella Miller Kids First Pediatric Research Program project <phs001138.v4.p2>, and were accessed from the Kids First Data Resource Portal (https://kidsfirstdrc.org/) and dbGaP (www.ncbi.nlm.nih.gov/gap). The PCGC Kids First study includes data sequenced by the Broad Institute (U24 HD090743-01). We acknowledge the use of CAVATICA as the cloud-based data analysis platform that supported this work. The authors also acknowledge OpenAI for editorial support.

Author contributions

Keyang Yu (Conceptualization [lead], Data Curation [lead], Formal Analysis [lead], Investigation [lead], Methodology [lead], Project Administration [lead], Resources [lead], Supervision [equal], Visualization [lead], Writing Original Draft [lead], Writing Review Editing [lead]), Haoquan Zhao (Conceptualization [equal], Methodology [equal], Resources [equal], Software [equal], Validation [equal], Writing Review Editing [supporting]), Andrea S. Wilderman (Conceptualization [supporting], Data Curation [supporting], Resources [supporting], Writing Original Draft [supporting], Writing Review Editing [supporting]), Tierra Farris (Data Curation [equal], Software [lead], Writing Review Editing [supporting]), Jessie Arce (Software [lead]), David Chen (Software [supporting]), Andrew R. Jackson (Software [supporting], Supervision [lead], Writing Review Editing [supporting]), Yiran Guo (Data Curation [supporting], Resources [supporting], Writing Original Draft [supporting]), Qi Li (Data Curation [supporting], Resources [supporting]), Bosko Jevtic (Software [equal], Supervision [lead]), Dubravka Jevtic (Software [equal], Supervision [equal]), Vuk Milinovic (Software [equal]), Yuankun Zhu (Software [supporting]), Jeremy Costanza (Software [supporting]), Eric Wenger (Project Administration [equal], Supervision [equal]), Chris Nemarich (Project Administration [equal], Supervision [equal]), Lisa Anderson (Project Administration [equal], Supervision [equal]), Aleksandar Mihajlović (Project Administration [equal], Supervision [lead]), Kristin Ardlie (Funding Acquisition [equal], Supervision [equal]), Shaine A. Morris (Data Curation [supporting], Writing Review Editing [supporting]), Matthew E. Roth (Project Administration [lead], Supervision [lead], Writing Review Editing [supporting]), Deanne M. Taylor (Data Curation [supporting], Funding Acquisition [equal], Supervision [equal]), Adam C. Resnick (Funding Acquisition [lead], Supervision [equal]), Lilei Zhang (Conceptualization [supporting], Data Curation [supporting], Supervision [supporting]), and Aleksandar Milosavljevic (Conceptualization [equal], Funding Acquisition [lead], Methodology [equal], Supervision [lead], Writing Original Draft [equal], Writing Review Editing [lead])

Supplementary material

Supplementary material is available at *Bioinformatics* online.

Conflict of interests

A.M. owns IP Genesis, Inc., a bioinformatics consulting business.

Funding

This work was supported by the National Institutes of Health Office of the Director [OT2 OD038054, U54 OD030165, OT2 OD030547]; the National Human Genome Research Institute [U24 HG009649, U24 HG012108]; and the National Institute on Aging [U01 AG072439].

Data availability

The code and free access data underlying the congenital heart disease use case are available in Zenodo at https://doi.org/10.5281/zenodo.17833070. Packages for CAR, GLR, CFDE LDH, and

Alias service of the Linked Data Hub are available under MIT license in genboree node repository at https://genboree.org/veraccio/#/. All data submitted to the above databases can be accessed through API endpoints. Sequence data underlying this article are available through the NIH database of Genotypes and Phenotypes (dbGaP) under accession [phs001138.v3.p2, phs001194.v3.p2, phs001194.v4.p3]. Access to individual-level data is controlled; qualified researchers may obtain access by submitting a dbGaP Data Access Request to the relevant Data Access Committee in accordance with NIH dbGaP policies and the dataset's data use limitations.

References

Ackoff RL. From data to wisdom. *J Appl Syst Anal* 1989;**16**:3–9.

Alharbi R, Ahmed U, Dobriy D *et al.* Exploring the role of generative AI in constructing knowledge graphs for drug indications with medical context: 15th international conference on semantic web applications and tools for health care and life sciences, SWAT4HCLS 2024. *CEUR Workshop Proc* 2024; **3890**:1–10.

Blake B, Brady LI, Rouse NA *et al.* The efficacy of whole genome sequencing and RNA-Seq in the diagnosis of whole exome sequencing negative patients with complex neurological phenotypes. *J Pediatr Genet* 2023;**12**:206–12.

Bonello J, Cachia E, Alfino N. AutoFAIR—a portal for automating FAIR assessments for bioinformatics resources. *Biochim Biophys Acta Gene Regul Mech* 2022;**1865**:194767.

Borchert F, Mock A, Tomczak A *et al.* Knowledge bases and software support for variant interpretation in precision oncology. *Brief Bioinform* 2021;**22**:bbab134.

Brush MH, Shefchek K, Haendel M. SEPIO: 2016 Joint International Conference on Biological Ontology and BioCreative—Food, Nutrition, Health and Environment for the 9 Billion, ICBO-BioCreative, 2016. In: *CEUR Workshop Proceedings*, Corvallis, OR, United States, 2016, 1747.

Charbonneau AL, Brady A, Czajkowski K *et al.* Making common fund data more findable: catalyzing a data ecosystem. *Gigascience* 2022;**11**:giac105.

Clark T, Caufield H, Parker JA *et al.* AI-readiness for biomedical data: Bridge2AI recommendations. bioRxiv, 2024.10.23.619844. 2024, preprint: not peer reviewed.

Clarke DJB, Wang L, Jones A *et al.* FAIRshake: toolkit to evaluate the FAIRness of research digital resources. *Cell Syst* 2019; **9**:417–21.

Coorens THH, Guillaumet-Adkins A, Kovner R *et al.*; dGTEx Consortium. The human and non-human primate developmental GTEx projects. *Nature* 2025;**637**:557–64.

Coppola CJ, Ramaker RC, Mendenhall EM. Identification and function of enhancers in the human genome. *Hum Mol Genet* 2016;**25**:R190–7.

Diehl AD, Meehan TF, Bradford YM *et al.* The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J Biomed Semantics* 2016;**7**:44.

Dunham I, Kundaje A, Aldred SF *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; **489**:57–74.

Emani PS, Liu JJ, Clarke D *et al.*; PsychENCODE Consortium. Single-cell genomics and regulatory networks for 388 human brains. *Science* 2024;**384**:eadi5199.

Garcia L, Bolleman J, Gehant S *et al.*; UniProt Consortium. FAIR adoption, assessment and challenges at UniProt. *Sci Data* 2019;**6**:175.

Goar W, Babb L, Chamala S *et al.* Development and application of a computable genotype model in the GA4GH variation representation specification. *Pac Symp Biocomput* 2023; **28**:383–94.

Griffith M, Spies NC, Krysiak K *et al.* CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat Genet* 2017;**49**:170–4.

Grigoriu A, Zaveri A, Weiss G *et al.* SIENA: semi-automatic semantic enhancement of datasets using concept recognition. *J Biomed Semantics* 2021;**12**:5.

Helle E, Pihkala J, Turunen R *et al.* Rare variants in genes associated with cardiomyopathy are not common in hypoplastic left heart syndrome patients with myocardial dysfunction. *Front Pediatr* 2020;**8**. https://doi.org/10.3389/fped.2020.596840

Hodder A, Leiter SM, Kennedy J *et al.* Benefits for children with suspected cancer from routine whole-genome sequencing. *Nat Med* 2024;**30**:1905–12.

Iascone M, Ciccone R, Galletti L *et al.* Identification of de novo mutations and rare variants in hypoplastic left heart syndrome. *Clin Genet* 2012;**81**:542–54.

Kapoor M, Ventura ES, Walsh A *et al.* Building a FAIR data ecosystem for incorporating single-cell transcriptomics data into agricultural genome to phenome research. *Front Genet* 2024;**15**: 1460351. https://doi.org/10.3389/fgene.2024.1460351

Kundaje A, Meuleman W, Ernst J *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 2015;**518**:317–30.

Landrum MJ, Chitipiralla S, Kaur K *et al.* ClinVar: updates to support classifications of both germline and somatic variants. *Nucleic Acids Res* 2025;**53**:D1313–21.

Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. *Cell* 2013;**152**:1237–51.

Levine M, Cattoglio C, Tjian R. Looping back to leap forward: transcription enters a new era. *Cell* 2014;**157**:13–25.

Liu S, Chen W, Zhan Y *et al.* DNAH11 variants and its association with congenital heart disease and heterotaxy syndrome. *Sci Rep* 2019;**9**:6683.

Luo Y, Hitz BC, Gabdank I *et al.* New developments on the encyclopedia of DNA elements (ENCODE) data portal. *Nucleic Acids Res* 2020;**48**:D882–9.

Ma XR, Conley SD, Kosicki M *et al.*; Birth Defects Research Laboratory. Molecular convergence of risk variants for congenital heart defects leveraging a regulatory map of the human fetal heart. medRxiv, 2024.11.20.24317557, 2024, preprint: not peer reviewed.

Mardis ER. A decade's perspective on DNA sequencing technology. *Nature* 2011;**470**:198–203.

Moore JE, Purcaro MJ, Pratt HE *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* 2020;**583**:699–710.

Morton SU, Quiat D, Seidman JG *et al.* Genomic frontiers in congenital heart disease. *Nat Rev Cardiol* 2022;**19**:26–42.

Mungall CJ, Torniai C, Gkoutos GV *et al.* Uberon, an integrative multi-species anatomy ontology. *Genome Biol* 2012;**13**:R5.

NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic Acids Res* 2013; **41**:D8–20.

Onuchic V, Lurie E, Carrero I *et al.* Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci. Science 2018;**361**:eaar3146. https://doi.org/10.1126/science.aar3146

Orlov YL, Chen M, Kolchanov NA *et al.* BGRS: bioinformatics of genome regulation and data integration. J Integr Bioinform 2023;**20**:20230032. https://doi.org/10.1515/jib-2023-0032

Oudelaar AM, Higgs DR. The relationship between genome structure and function. *Nat Rev Genet* 2021;**22**:154–68.

Pawliczek P, Patel RY, Ashmore LR *et al.*; Clinical Genome (ClinGen) Resource. ClinGen Allele Registry links information about genetic variants. *Hum Mutat* 2018;**39**:1690–701.

Perez G, Barber GP, Benet-Pages A *et al.* The UCSC Genome Browser database: 2025 update. *Nucleic Acids Res* 2025; **53**:D1243–9.

Phillips KA, Trosman JR, Douglas MP *et al.* US private payers' perspectives on insurance coverage for genome sequencing versus exome sequencing: a study by the Clinical Sequencing Evidence-Generating Research Consortium (CSER). *Genet Med* 2022;**24**:238–44.

Preissl S, Gaulton KJ, Ren B. Characterizing cis-regulatory elements using single-cell epigenomics. *Nat Rev Genet* 2023; **24**:21–43.

Richter F, Morton SU, Kim SW *et al.* Genomic analyses implicate noncoding de novo variants in congenital heart disease. *Nat Genet* 2020;**52**:769–77.

Rozenblatt-Rosen O, Stubbington MJT, Regev A *et al.* The Human Cell Atlas: from vision to reality. *Nature* 2017; **550**:451–3.

Rozowsky J, Gao J, Borsari B *et al.* The EN-TEx resource of multitissue personal epigenomes & variant-impact models. *Cell* 2023;**186**:1493–511.e40.

Seal RL, Braschi B, Gray K *et al.* Genenames.org: the HGNC resources in 2023. *Nucleic Acids Res* 2023;**51**:D1003–9.

Smemo S, Campos LC, Moskowitz IP *et al.* Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *Hum Mol Genet* 2012;**21**:3255–63.

Sun HY. Prenatal diagnosis of congenital heart defects: echocardiography. *Transl Pediatr* 2021;**10**:2210–24.

The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 2020; **369**:1318–30.

Theis JL, Hu JJ, Sundsbak RS *et al.* Genetic association between hypoplastic left heart syndrome and cardiomyopathies. *Circ: Genomic and Precision Medicine* 2021;**14**. https://doi.org/10.1161/CIRCGEN.120.003126

Theis JL, Olson TM. Whole genome sequencing in hypoplastic left heart syndrome. *J Cardiovasc Dev Dis* 2022;**9**:117.

Thorogood A, Rehm HL, Goodhand P *et al.* International federation of genomic medicine databases using GA4GH standards. Cell Genomics 2021;**1**:100032. https://doi.org/10.1016/j.xgen.2021.100032

Trosman JR, Douglas MP, Liang S-Y *et al.* Insights from a temporal assessment of increases in US private payer coverage of tumor sequencing from 2015 to 2019. *Value Health* 2020; **23**:551–8.

VanOudenhove J, Yankee TN, Wilderman A *et al.* Epigenomic and transcriptomic dynamics during human heart organogenesis. *Circ Res* 2020;**127**:e184–209.

Wilkinson MD, Dumontier M, Aalbersberg I *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;**3**:160018.

Wilkinson MD, Sansone S-A, Schultes E *et al.* A design framework and exemplar metrics for FAIRness. *Sci Data* 2018;**5**:180118.

Xin J, Mark A, Afrasiabi C *et al.* High-performance web services for querying gene and variant annotation. *Genome Biol* 2016; **17**:91.

Yagi H, Liu X, Gabriel GC *et al.* The genetic landscape of hypoplastic left heart syndrome. *Pediatr Cardiol* 2018;**39**:1069–81.

Yates AD, Achuthan P, Akanni W *et al.* Ensembl 2020. *Nucleic Acids Res* 2020;**48**:D682–8.

Yates AD, Adams J, Chaturvedi S *et al.* Refget: standardized access to reference sequences. *Bioinformatics* 2021;**38**:299–300.

Zaidi S, Brueckner M. Genetics and genomics of congenital heart disease. *Circ Res* 2017;**120**:923–40.

Zaveri A, Dastgheib S, Wu C *et al.* smartAPI: towards a more intelligent network of web APIs. In: Blomqvist E, Maynard D, Gangemi A (eds.), *The Semantic Web*. Cham: Springer International Publishing, 2017, 154–69.

Zhou J, Jiang J, Han Z *et al.* Streamline automated biomedical discoveries with agentic bioinformatics. *Brief Bioinform* 2025; **26**:bbaf505.

Zhou W-Z, Li W, Shen H *et al.* CHDbase: a comprehensive knowledgebase for congenital heart disease-related genes and clinical manifestations. *Genom Proteom Bioinform* 2023; **21**:216–27.

Bioinformatics, 2026, 42, 1–12
https://doi.org/10.1093/bioinformatics/btag013
Original Paper