Review

CD2K: An Accessible Framework for Leveraging Public Omics Data in Biomedical Research Training

Darawan Rinchai , Mathieu Garand , Mohammed Toufiq , Basirudeen Syed Ahamed Kabeer , Nico Marr , Damien Chaussabel *

Posted Date: 11 October 2024

doi: 10.20944/preprints202410.0933.v1

Keywords: Omics data; Biomedical research training; Data reuse; Low- and middle-income countries (LMICs)

Review

CD2K: An Accessible Framework for Leveraging Public Omics Data in Biomedical Research Training

Darawan Rinchai [1], Mathieu Garand [2], Mohammed Toufiq [3], Basirudeen Syed Ahamed Kabeer [4], Nico Marr [5] and Damien Chaussabel [3,*]

[1] St Jude Research Hospital, Memphis, TN, USA
[2] Division of Pediatric Cardiothoracic Surgery, Department of Surgery, Washington University School of Medicine, St. Louis, MO, USA
[3] The Jackson Laboratory, Farmington, CT, USA
[4] Sidra Medicine, Doha, Qatar
[5] College of Health and Life Sciences, Hamad Bin Khalifa University, Doha, Qatar
* Correspondence: to: Damien.chaussabel@jax.org

Abstract: The exponential growth of publicly available omics data presents both an opportunity and a challenge for biomedical researchers, particularly those in low- and middle-income countries (LMICs). The Collective Omics Data to Knowledge (CD2K) initiative aims to address this challenge by providing an accessible framework for biomedical research training. This review describes the development and implementation of the CD2K program, which comprises three core modules: COD1 (reductionist interpretation of collective omics data), COD2 (creation of curated dataset collections), and COD3 (re-analysis of omics data on a global scale). The CD2K approach emphasizes the reuse and reinterpretation of public data, integrating literature mining and emerging technologies like Large Language Models (LLMs). A key feature of the program is its focus on accessibility, designed to make the exploitation of large-scale datasets amenable to researchers without extensive data science skills. The curriculum aims to equip trainees with a range of skills, from basic data interpretation to more advanced bioinformatics analysis, with an emphasis on producing tangible outputs such as peer-reviewed publications, which directly address career development needs. The CD2K initiative has involved researchers from multiple institutions across several countries, resulting in several publications and publicly available dataset collections. While still in its early stages, the program shows promise in providing a structured framework for leveraging public omics data in biomedical research. This review also discusses the current limitations of the CD2K approach and ongoing efforts to expand its reach. By offering an accessible model for building research capacity, the CD2K initiative represents a step towards fostering data-driven discovery in global biomedical research, particularly in resource-limited settings.

Keywords: omics data; biomedical research training; data reuse; low- and middle-income countries (LMICs)

1. Executive Summary/Introduction

The Collective Omics Data to Knowledge (CD2K) initiative represents an original approach to biomedical research training, leveraging the vast potential of publicly available omics datasets [1]. This program primarily aims to address a critical need in the field: supporting the development of research capacity, particularly in low- and middle-income countries (LMICs). As a byproduct of these training activities, the program also contributes to generating actionable knowledge from the wealth of existing data [1,2].

At its core, CD2K is built on the principle that the endpoint of the training curriculum should be publication, aligning with the primary currency for advancing research careers [1]. This focus on generating "actionable" or translational knowledge ensures that participants not only develop crucial skills but also produce tangible outputs with real-world impact, serving as a source of motivation for trainees [2].

The CD2K initiative encompasses three core training curricula [1]:

 doi:10.20944/preprints202410.0933.v1

1. COD1: Focuses on reductionist interpretation of collective omics data
2. COD2: Involves the creation of curated dataset collections
3. COD3: Trains participants in the re-analysis of omics data on a global scale

These modules equip researchers with the skills to navigate public data repositories, identify knowledge gaps, apply novel analytical approaches, and ultimately generate publishable findings [1,2]. Importantly, some of these programs, particularly the COD1 workflows, are accessible to participants who do not have extensive data science training or experience. This accessibility demonstrates that publication based on collections of public omics datasets and resulting career advancement are attainable for mainstream bench scientists [1].

As will be detailed further in this article, the success of the CD2K initiative is evidenced by numerous proof-of-principle publications from researchers who have completed one or more of the COD training modules. These successes demonstrate the feasibility of using collective omics data as a springboard for research output and career advancement, even in resource-limited settings [1,2].

Recent advancements in the program include the integration of Large Language Models (LLMs) into the COD1 workflow, further enhancing the efficiency and depth of analysis possible within this framework [2]. This integration represents a significant step forward in leveraging cutting-edge AI technologies for biomedical research and training.

As the CD2K initiative continues to evolve, it holds significant potential to democratize access to omics data and empower researchers, particularly in LMICs, to make meaningful contributions to biomedical knowledge [1,2]. By providing a structured framework for data reuse and emphasizing actionable outcomes, CD2K aligns with broader efforts to promote equity and accelerate discovery in global health research.

2. Historical Development and Overview of COD Training Modules

2.1. COD1: Reductionist Interpretation of Collective Omics Data

The COD1 module focuses on the reductionist interpretation of collective omics data, emphasizing gene-centric investigations. This approach has proven to be accessible to participants without extensive data science training, making it particularly valuable for mainstream bench scientists. The methodology involves selecting genes presenting differences in transcript abundance between study groups, assessing knowledge gaps by examining the overlap between gene-associated literature and study-relevant concepts, and validating observations using independent datasets when possible [1].

Since its inception, the COD1 training program has involved 49 unique individuals from 12 different institutions across 8 countries, demonstrating its global reach and collaborative nature. These workshops have resulted in 10 peer-reviewed publications, each focusing on a different gene and its potential role in various immunological and disease contexts (Table 1).

Table 1.

Gene	Number of Participants	Affiliations	Theme	Citation (PMID)
ALAS2	7	The Jackson Laboratory, The Rockefeller University, INSERM, Sidra Medicine, University of Bretagne Occidentale, CHU de Brest	Erythroid cells	37845713
ACVR1B	5	Mahidol University, Sidra Medicine, University of Washington	Sepsis	37020544
NUDT16	6	Sidra Medicine, Washington University School of Medicine	Sepsis	35174610
ERLIN1	6	Sidra Medicine	Sepsis	34439987
CST7	4	University of Sydney, Sidra Medicine	Acute inflammation	33868254
BANK1	6	University of Nantes, Sidra Medicine	B cells, tolerance	33815360

 doi:10.20944/preprints202410.0933.v1

ANXA3	10	Sidra Medicine	Sepsis	32682335
ACSL1	11	Sidra Medicine, The University of Texas MD Anderson Cancer Center	Sepsis	31681299
TKT	9	Mahidol University, Sidra Medicine, National Institute of Infectious Diseases (Japan), University of Western Australia	Inflammation	31415630
ADAM9	5	Khon Kaen University, Sidra Medicine	Tissue damage	27990250

The genes investigated through this program span a wide range of biological processes and disease states. For instance, ALAS2's role in erythroid cells [2], ACVR1B, NUDT16, ERLIN1, ANXA3, and ACSL1 in sepsis [3–7], CST7 in acute inflammation [8], BANK1 in B cell tolerance [9], TKT in inflammation [10], and ADAM9 in tissue damage [11] have all been explored.

While primarily serving as a training tool, these investigations have yielded potentially significant findings. For example, the study on ANXA3 suggested its potential role in neutrophil function during sepsis, which could have implications for sepsis management [6]. The investigation of ACSL1 pointed to its possible involvement in inflammasome activation in neutrophils during sepsis, with potential prognostic value [7].

It is important to note that while these findings suggest potential translational significance, they should be considered preliminary and require further validation. Nonetheless, they demonstrate the capacity of this training approach to generate hypotheses and identify potential biomarkers or therapeutic targets worthy of further investigation. Taken together this iteration of the COD1 module has successfully equipped participants with skills in navigating public data repositories, identifying knowledge gaps in biomedical literature, developing advanced PubMed query skills, and formulating and testing hypotheses.

The COD1 module has recently undergone significant evolution, with a sharpened focus on the development of targeted assays as an actionable endpoint (Figure 1). This shift aligns with the CD2K/COD program's emphasis on producing tangible, translatable outcomes alongside publications. The integration of Large Language Models (LLMs) into the workflow has markedly enhanced both the efficiency and depth of analysis possible within this framework [2]. This technological advancement has enabled the implementation of sophisticated workflows for gene prioritization and selection, followed by in-depth characterization of selected candidates, exemplified by studies such as the CEACAM6 analysis [52]. The new orientation of COD1 reflects a more holistic approach to biomarker discovery and validation, bridging the gap between high-throughput data analysis and practical clinical applications. To date, three workshops implementing this updated curriculum have been conducted The Jackson Laboratory in the US, Mahidol University in Thailand, and Khon Kaen University in Thailand. These workshops have provided participants with hands-on experience in leveraging cutting-edge AI technologies for biomedical research. This evolution of the COD1 module represents a significant step forward in translating complex genomic data into potential diagnostic and therapeutic tools, further enhancing the impact and relevance of the CD2K initiative in the field of precision medicine.

 doi:10.20944/preprints202410.0933.v1

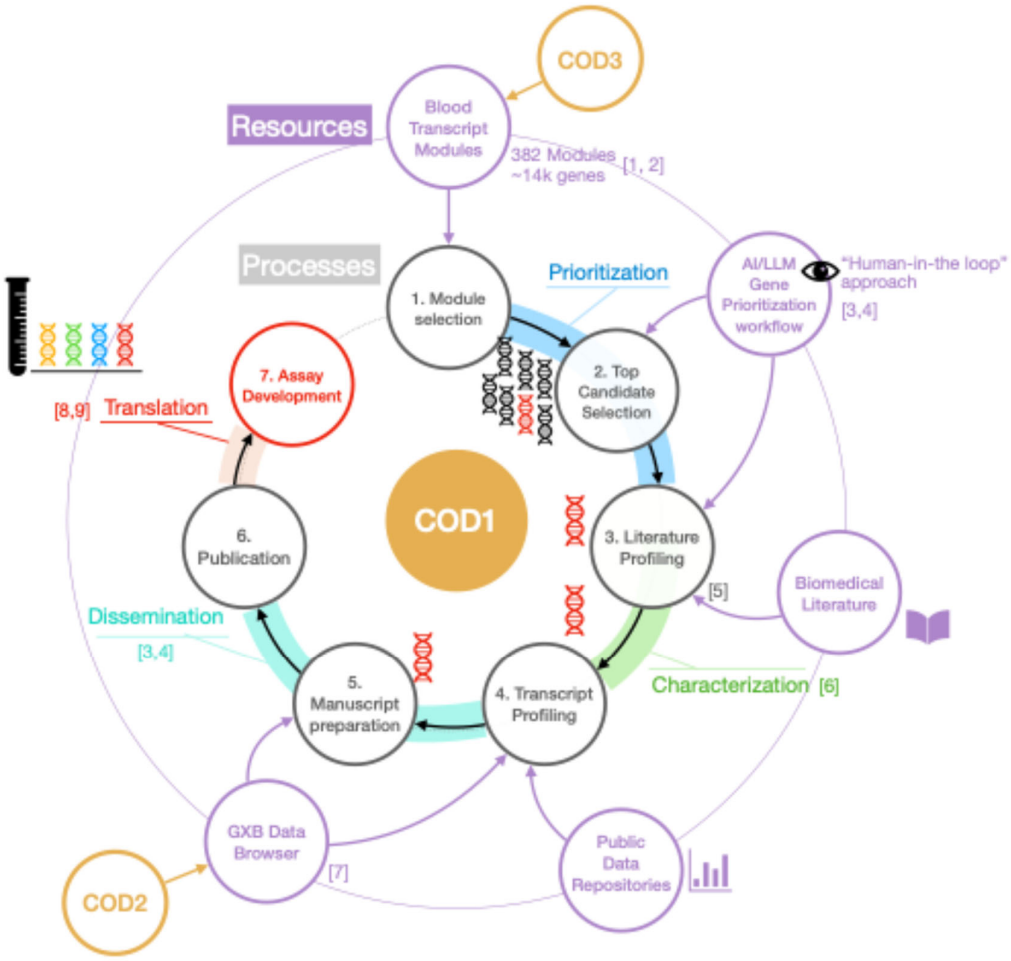

Figure 1: Workflow of the COD1 (Collective Omics Data 1) training module within the CD2K initiative. This diagram outlines the key processes and resources involved in an illustrative reductionist data-to-knowledge COD1 workflow. It integrates various resources (purple circles) such as: 1) the fixed BloodGen3 module repertoire - 382 modules, comprising over 14,000 genes, identified based on co-expression patterns, 2) custom AI/Large Language Models (LLM) prioritization workflows, including intermediate fact checking steps (human-in-the-loop approach), 3) the biomedical literature and 4) Public Omics Data Repositories (e.g. NCBI's Gene Expression Omnibus, GEO). The main processes (grey circles) include: 1) Module Selection, 2) Gene Prioritization - i.e. selecting a top gene candidate (red helix) among a pool formed by the genes comprised in the module of interest (in black), 3) Literature profiling, 4) Public transcriptome data profiling, 5) manuscript preparation, and 6) dissemination (posting of the manuscript on pre-print server, oral and/or poster presentations). Colored arrows indicate different stages of the process: Prioritization (blue), Characterization (green), Translation (red), and Dissemination (teal).[SEP][SEP]. References [1-9]: 1: Altman MC, Rinchai D, [...] and Chaussabel D. Development of a fixed module repertoire for the analysis and interpretation of blood transcriptome data. Nat Commun. 2021 Jul 19;12(1):4385. 2: Chaussabel D, Baldwin N. Democratizing systems immunology with modular transcriptional repertoire analyses. Nat Rev Immunol. 2014 Apr;14(4):271-80. 3: Subba B, Toufiq M, [...] and Chaussabel D. Human-augmented large language model-driven selection of glutathione peroxidase 4 as a candidate blood transcriptional biomarker for circulating erythroid cells. Sci Rep. 2024 Oct 5;14(1):23225. 4: Toufiq M, Rinchai D, [...] and Chaussabel D. Harnessing large language models (LLMs) for candidate gene prioritization and selection. J Transl Med. 2023 Oct 16;21(1):728. 5: Al Ali F, Marr AK, [...] and Chaussabel D. Organizing training workshops on gene literature retrieval, profiling,

 doi:10.20944/preprints202410.0933.v1

and visualization for early career researchers. F1000Res. 2023 May 11;10:275. 6: Rinchai D, Chaussabel D. Assessing the potential relevance of CEACAM6 as a blood transcriptional biomarker. F1000Res. 2024 Apr 4;11:1294. 7: Speake C, Presnell S, [...] and Chaussabel D. An interactive web application for the dissemination of human systems immunology data. J Transl Med. 2015 Jun 19;13:196. 8: Brummaier T, Rinchai D, [...] and Chaussabel D. Design of a targeted blood transcriptional panel for monitoring immunological changes accompanying pregnancy. Front Immunol. 2024 Jan 30;15:1319949. 9: Rinchai D, Syed Ahamed Kabeer B, [...] and Chaussabel D. A modular framework for the development of targeted Covid-19 blood transcript profiling panels. J Transl Med. 2020 Jul 31;18(1):291.

COD2: Creation of Curated Collective Omics Dataset Collections

The COD2 module focuses on the creation of curated dataset collections, emphasizing the organization and accessibility of large-scale omics data (Figure 2). This approach enables researchers to leverage existing public data repositories effectively, promoting data reuse and fostering new discoveries [12].

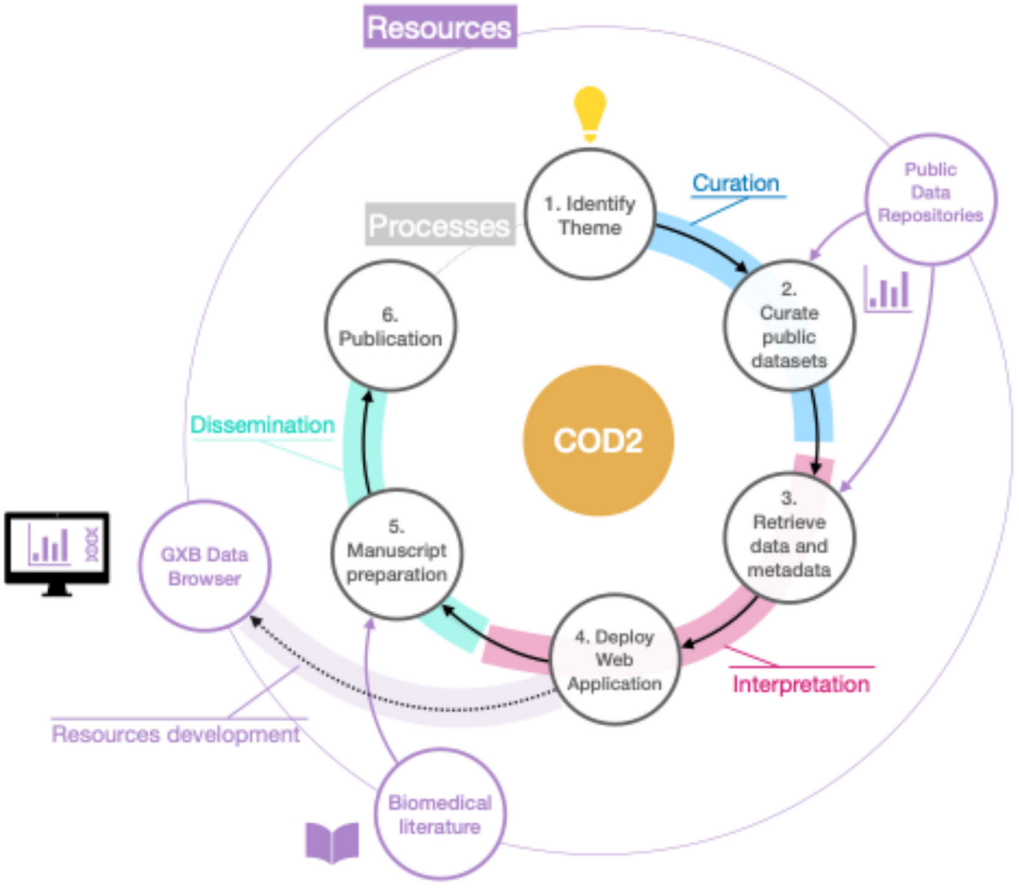

Figure 2. Workflow of the COD2 (Collective Omics Data 2) training module within the CD2K initiative. This diagram illustrates the COD2 workflow for enhancing the accessibility of omics data. Resources (purple circles) include Public Data Repositories and Biomedical Literature. The workflow comprises six main processes (grey circles): 1) Identify Theme, 2) Curate Public Datasets, 3) Retrieve Data and Metadata, 4) Deploy Web Application, 5) Manuscript Preparation, and 6) Publication. Key aspects are highlighted by colored arrows: Curation (blue), Interpretation (pink), and Dissemination (teal). The GXB Data Browser, developed as part of the Resources Development process (purple

 doi:10.20944/preprints202410.0933.v1

dotted arrow), is a key output that facilitates data exploration. This workflow trains researchers in creating resources that democratize access to complex omics data, supporting data reuse and reproducibility in biomedical research.

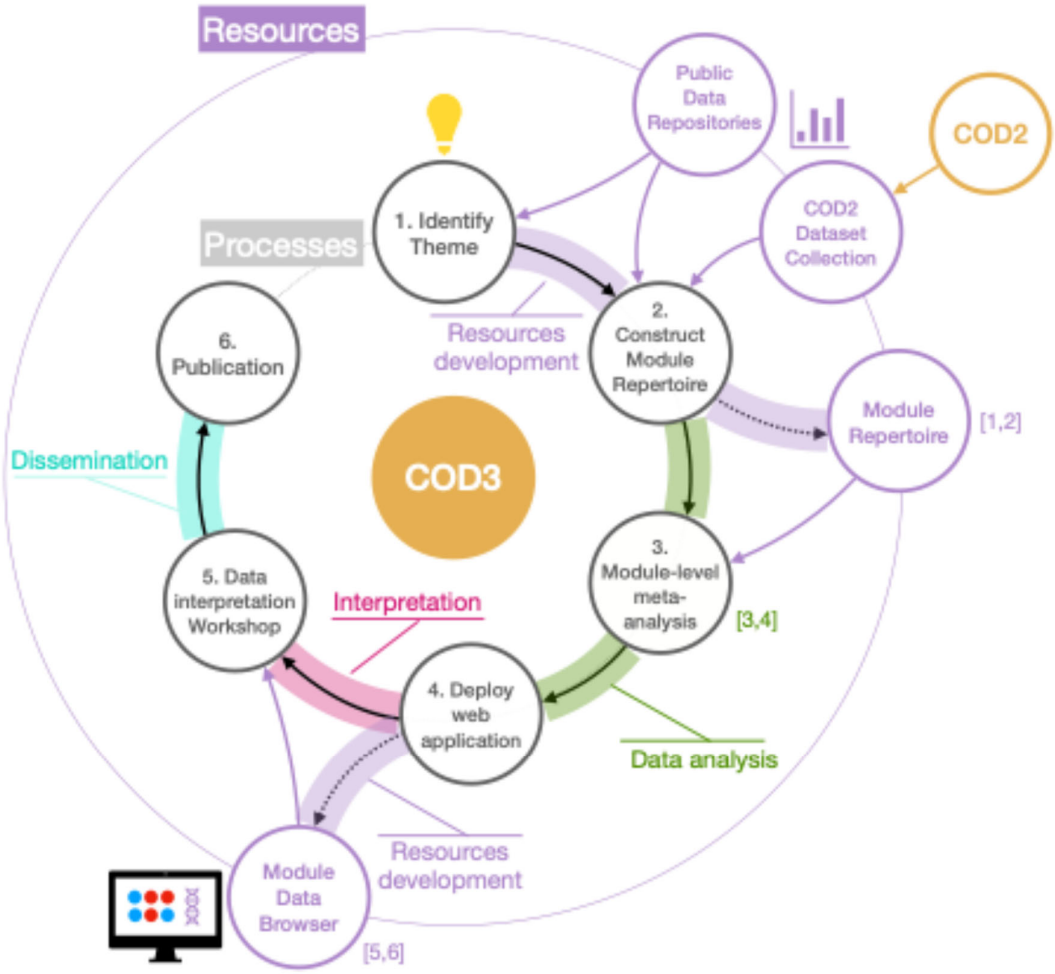

Figure 3. Workflow of the COD3 (Collective Omics Data 3) training module within the CD2K initiative. This diagram illustrates the COD3 workflow for large-scale omics data analysis and interpretation. Resources (purple circles) include Public Data Repositories, COD2 Dataset Collections. The workflow comprises six main processes (grey circles): 1) Identify Theme, 2) Construct a Module Repertoire, 3) Module-level Meta-analysis, 4) Deploy Web Application, 5) Data Interpretation Workshop, and 6) Publication. Key aspects are highlighted by colored arrows: Resources Development (purple), Data Analysis (green), Interpretation (pink), and Dissemination (teal). The Module Data Browser, developed as part of the Resources Development process, facilitates data exploration and interpretation. This workflow trains researchers in advanced data analysis techniques while making complex omics data more accessible to those without extensive computational backgrounds, supporting data-driven discovery in biomedical research. References [1-6]: 1: Altman MC, Rinchai D, [...] and Chaussabel D. Development of a fixed module repertoire for the analysis and interpretation of blood transcriptome data. Nat Commun. 2021 Jul 19;12(1):4385. doi: 10.1038/s41467-021-24584-w. PMID: 34282143; PMCID: PMC8289976. 2: Chaussabel D, Baldwin N. Democratizing systems immunology with modular transcriptional repertoire analyses. Nat Rev Immunol. 2014 Apr;14(4):271-80. doi: 10.1038/nri3642. PMID: 24662387; PMCID: PMC4118927. 3: Rinchai D, Altman MC, [...]

 doi:10.20944/preprints202410.0933.v1

and Chaussabel D. Definition of erythroid cell-positive blood transcriptome phenotypes associated with severe respiratory syncytial virus infection. Clin Transl Med. 2020 Dec;10(8):e244. doi: 10.1002/ctm2.244. PMID: 33377660; PMCID: PMC7733317. 4: Rawat A, Rinchai D, [...] and Chaussabel D. A Neutrophil-Driven Inflammatory Signature Characterizes the Blood Transcriptome Fingerprint of Psoriasis. Front Immunol. 2020 Nov 24;11:587946. doi: 10.3389/fimmu.2020.587946. PMID: 33329570; PMCID: PMC7732684. 5: Rinchai D, Brummaier T, [...] and Chaussabel D. A data browsing application for accessing gene and module-level blood transcriptome profiles of healthy pregnant women from high- and low-resource settings. Database (Oxford). 2024 Apr 2;2024:baae021. doi: 10.1093/database/baae021. PMID: 38564425; PMCID: PMC10986794. 6: Bettacchioli E, Chiche L, [...] and Rinchai D. An interactive web application for exploring systemic lupus erythematosus blood transcriptomic diversity. Database (Oxford). 2024 May 28;2024:baae045. doi: 10.1093/database/baae045. PMID: 38805754; PMCID: PMC11131423.

The methodology involves identifying relevant datasets from public repositories such as the NCBI Gene Expression Omnibus (GEO), curating and annotating these datasets, and making them accessible through user-friendly web applications [13]. These applications, such as the Gene Expression Browser (GXB), allow for interactive query and visualization of integrated large-scale data [14].

Since its inception, the COD2 training program has involved 84 unique individuals from over 20 different institutions across multiple countries, demonstrating its global reach and collaborative nature. These efforts have resulted in 12 publicly available dataset collections, each focusing on a different biological theme or disease context (Table 2).

Table 2.

Participants	Theme	Datasets	Unique Profiles	PMID	Web Link
6	Systemic Lupus Erythematosus	1	157	38805754	https://immunology-research.shinyapps.io/LUPUCE/
11	Pregnancy	2+	15+	38564425	https://thejacksonlaboratory.shinyapps.io/BloodGen3_Pregnancy/
4	Primary Immunodeficiencies	18	Not specified	31559014	http://pid.gxbsidra.org/dm3/geneBrowser/list
6	IgE-mediated Atopic Diseases	33	1860	31290545	http://ige.gxbsidra.org/dm3/geneBrowser/list
4	Viral Respiratory Tract Infection and Vaccination	31	6648	31231515	http://vri1.gxbsidra.org/dm3/geneBrowser/list
18	Breast Cancer	13	2142	29527288	http://breastcancer.gxbsidra.org/dm3/geneBrowser/list
4	In Vitro Fertilization and Polycystic Ovary Syndrome	12	85	28413616	http://ivf.gxbsidra.org/dm3/landing.gsp
7	Hematopoietic Cells in Early Life	32	2129	27347375	http://developmentalimmunology.gxbsidra.org/dm3/geneBrowser/list
6	Placenta Development and Pathologies	24	759	27303626	http://placentalendocrinology.gxbsidra.org/dm3/landing.gsp

 doi:10.20944/preprints202410.0933.v1

5	Human Monocyte Immunobiology	93	4516	27158452	http://monocyte.gxbsidra.org/dm3/landing.gsp
5	HIV Infection	34	2717	27134731	http://hiv.gxbsidra.org/dm3/geneBrowser/list
8	Systemic Inflammatory Responses to Sepsis	62	5719	34663591	http://sepsis.gxbsidra.org/dm3/geneBrowser/list

The dataset collections span a wide range of biological processes and disease states, including systemic lupus erythematosus [15], pregnancy [16], primary immunodeficiencies [17], IgE-mediated atopic diseases [18], viral respiratory tract infections [19], breast cancer [20], in vitro fertilization [21], hematopoietic cells in early life [22], placental development [23], monocyte immunobiology [24], HIV infection [25], and sepsis [26].

While primarily serving as a training tool and resource for the research community, these curated dataset collections have yielded significant benefits. They have facilitated the re-analysis of public data, leading to new insights and hypotheses. For example, the breast cancer dataset collection has been used to investigate immunologic classifications of tumors [20], while the sepsis collection has enabled the identification of specific human blood gene signatures in response to infection [26].

The COD2 module has successfully equipped participants with skills in data curation, annotation, and the development of user-friendly interfaces for data exploration. Recent advancements include the integration of module-level data browsing capabilities [15,16], which ties in with COD3 activities and enhances the depth of analysis possible within this framework.

The impact of the COD2 module extends beyond the immediate training outcomes. By making curated dataset collections publicly available, it has contributed to the broader scientific community's efforts in promoting data reuse and reproducibility in biomedical research. These resources serve as valuable tools for hypothesis generation, validation of findings across multiple studies, and the identification of robust biomarkers or therapeutic targets.

COD3: Re-Analysis of Collective Omics Data on a Global Scale

The COD3 module focuses on the analysis and interpretation of large-scale profiling data, particularly aimed at individuals with data science training or those interested in pursuing a career in this field. However, it is worth noting that data science expertise is not always a prerequisite for participation in COD3 activities.

Central to the COD3 approach is its reliance on well-characterized fixed module repertoires, such as the BloodGen3 module repertoire [27], a fixed set of 382 transcriptional modules encompassing 14,168 transcripts. This repertoire, along with the accompanying R package BloodGen3Module [28], and web applications for module-level data browsing [29,30], form the core resources leveraged in COD3 training activities (**Figure 3**).

The COD3 training activities have explored multiple avenues, as evidenced by several published use cases (Table 3):

Table 3.

PMID	Title	Participants	Theme	Resource /Training	Key Aspects
34282143	Development of a fixed module repertoire for the analysis and interpretation of blood transcriptome data	37	Development of BloodGen3 module repertoire	Resource	Fixed module repertoire for blood transcriptome analysis

 doi:10.20944/preprints202410.0933.v1

33624743	BloodGen3Module: blood transcriptional module repertoire analysis and visualization using R	7	Development of R package for BloodGen3	Resource	R package for module-level analysis and visualization
33329570	A Neutrophil-Driven Inflammatory Signature Characterizes the Blood Transcriptome Fingerprint of Psoriasis	13	Meta-analysis of psoriasis blood transcriptome	Training	Module-level meta-analysis for pathogenesis investigation
33377660	Definition of erythroid cell-positive blood transcriptome phenotypes associated with severe respiratory syncytial virus infection	14	Meta-analysis of RSV blood transcriptome	Training	Module-level meta-analysis for biomarker discovery
38352867	Design of a targeted blood transcriptional panel for monitoring immunological changes accompanying pregnancy	14	Development of pregnancy-specific blood transcriptome panel	Training	Data and knowledge-driven biomarker discovery
32736569	A modular framework for the development of targeted Covid-19 blood transcript profiling panels	18	Development of COVID-19 blood transcriptome panels	Training	Data and knowledge-driven biomarker discovery
38805754	An interactive web application for exploring systemic lupus erythematosus blood transcriptomic diversity	6	SLE blood transcriptome data browsing application	Resource	Web application for module-level data browsing
38564425	A data browsing application for accessing gene and module-level blood transcriptome profiles of healthy pregnant women from high- and low-resource settings	11	Pregnancy blood transcriptome data browsing application	Resource	Web application for module-level data browsing

1. Meta-analyses at the module level: These studies encompass multiple independent datasets, focusing either on investigating disease pathogenesis or biomarker discovery. For example, a study on psoriasis [31] utilized module-level meta-analysis to investigate the neutrophil-driven inflammatory signature characterizing the blood transcriptome fingerprint of the disease. Similarly, research on respiratory syncytial virus (RSV) infection [32] employed module-level

 doi:10.20944/preprints202410.0933.v1

meta-analysis for biomarker discovery, identifying erythroid cell-positive blood transcriptome phenotypes associated with severe RSV infection.

2. Data and knowledge-driven biomarker discovery: Several studies have employed this approach, including work on pregnancy [33], COVID-19 [34], and most recently, the use of Large Language Models (LLMs) for developing a generic immune profiling assay [35]. This last study is particularly noteworthy as it demonstrates the potential of integrating artificial intelligence into the biomarker discovery process.
3. Data interpretation workshops: These workshops, supported by BloodGen3 applications, involve participants with expertise in medicine and immunology who may not have extensive data science backgrounds. These activities emphasize the accessibility of COD3 workflows, even when focusing on the analysis of large-scale datasets. A manuscript describing this approach is currently in preparation.

The LLM-enabled workflow, detailed in the recent publication by Toufiq et al. [35], demonstrates how generative AI technology can be harnessed for knowledge-driven candidate gene prioritization and selection. By leveraging both data-driven and knowledge-based approaches, this integrated COD3 workflow represents a means to significantly improve candidate gene prioritization, enabling, for instance, the development of targeted profiling assays.

The resources developed as part of COD3, particularly the BloodGen3 module repertoire [27] and the BloodGen3Module R package [28], have facilitated the creation of interactive web applications for exploring blood transcriptomic diversity in various conditions. Notable examples include applications for systemic lupus erythematosus [29] and pregnancy [30], which allow researchers to interactively explore module-level blood transcriptome data.

Overall, the COD3 module has successfully equipped participants with advanced skills in large-scale data analysis, interpretation, and visualization. The integration of novel technologies like LLMs and the development of user-friendly web applications have further enhanced the accessibility and impact of these training activities.

Discussion

The Collective Omics Data to Knowledge (CD2K) initiative represents a novel approach to biomedical research training, leveraging the vast potential of publicly available omics datasets. Through its three core modules - COD1, COD2, and COD3 - the program addresses critical needs in the field: supporting research capacity development, particularly in low- and middle-income countries (LMICs), and generating actionable knowledge from existing data [36]. This approach aligns with growing global efforts to enhance data science capabilities in biomedicine and promote data-driven discovery [37,38].

In recent years, several initiatives have emerged to address the need for data science training in biomedical research. The NIH Big Data to Knowledge (BD2K) program, launched in 2012, has been at the forefront of these efforts, funding numerous training programs and resources to enhance biomedical big data utilization [39]. Similarly, the European Life-Science Infrastructure for Biological Information (ELIXIR) has developed extensive bioinformatics training programs across Europe [40]. In LMICs, networks like H3ABioNet have been instrumental in building bioinformatics capacity across Africa [41].

However, many of these programs have primarily focused on developing computational skills and infrastructure [42]. While essential, this approach may not fully address the needs of researchers who lack extensive programming backgrounds or those working in resource-limited settings. Furthermore, the translation of big data skills into actionable biomedical knowledge and career advancement opportunities remains a challenge [43].

The CD2K initiative builds upon these existing efforts while introducing novel elements to address some of these gaps. By emphasizing the reuse and reinterpretation of public data, CD2K provides a cost-effective means for researchers, particularly in LMICs, to contribute meaningfully to global scientific discourse. The program's focus on producing tangible outputs, such as peer-

 doi:10.20944/preprints202410.0933.v1

reviewed publications, directly addresses the critical aspect of career development often overlooked in short-term training initiatives [44].

The CD2K initiative stands out in several key aspects, addressing some of the limitations observed in traditional data science training programs:

1. Interdisciplinary Curriculum: The CD2K program integrates key concepts from genomics, bioinformatics, and immunology, as well as knowledge discovery and dissemination. This interdisciplinary approach provides trainees with a comprehensive understanding of how to translate omics data into actionable biomedical insights [45]. Unlike programs that focus solely on computational skills, CD2K emphasizes the biological context and translational potential of data analysis.
2. Advanced Technology Integration: The recent incorporation of Large Language Models (LLMs) into the curriculum, particularly in the COD1 module, represents a significant advancement in biomedical data analysis training [35]. LLMs offer powerful tools for literature mining, hypothesis generation, and even candidate gene prioritization. By introducing trainees to these cutting-edge technologies, CD2K prepares them for the future of data-driven biomedical research.
3. Publication as a Training Endpoint: A unique feature of CD2K is its emphasis on publication as a training outcome. This approach aligns practical training with tangible academic contributions, addressing a critical need for career advancement in research [46]. It provides trainees, especially those from LMICs, with opportunities to build their publication record and contribute to the global scientific discourse.
4. Hands-On Training with Real-World Datasets: CD2K offers hands-on experience with actual research datasets, enhancing the practical learning experience. This approach bridges the gap between theoretical knowledge and real-world application, a challenge often faced in traditional bioinformatics training programs [47].
5. Enhancing Data-to-Knowledge Translation: The program builds on traditional data science skills by teaching how to interpret complex datasets within broader scientific contexts. This focus on knowledge translation is crucial for developing actionable biomedical insights [48].
6. Accessibility and Inclusivity: While the program includes advanced data analysis techniques, it is designed to be accessible to researchers without extensive computational backgrounds. This inclusivity is particularly valuable for engaging a broader range of biomedical researchers, including those from LMICs or with primarily wet-lab experience [49].

Limitations and Future Directions

While the CD2K initiative has demonstrated significant success, it is important to acknowledge its limitations. First, the program's reliance on publicly available datasets may limit its applicability in some specialized research areas where data sharing is restricted. Second, the emphasis on publication as an endpoint, while valuable for career development, may inadvertently prioritize "publishable" findings over other important but less immediately impactful research outcomes.

Despite these limitations, ongoing efforts are being made to expand the reach and impact of the CD2K approach. A key development in this direction is the creation of workshop toolkits designed to disseminate the CD2K methodology more widely. These toolkits aim to equip early-career researchers with essential skills for navigating the increasingly data-rich landscape of biomedical research.

One such toolkit focuses on gene literature retrieval, profiling, and visualization [50]. This hands-on curriculum teaches participants how to build and refine PubMed queries, identify key concepts in the literature, measure concept prevalence, and extract and present relevant information. The workshop uses the ISG15 gene as an illustrative case study, providing participants with practical experience in literature mining and data presentation.

Another toolkit presents a comprehensive workflow for retrieving, structuring, and aggregating information from both literature and large-scale data repositories [51]. This curriculum guides trainees through a stepwise process, from selecting a candidate gene to drafting and publishing a manuscript. It emphasizes the integration of literature-based knowledge with insights derived from public transcriptome datasets, providing a holistic approach to biomarker assessment. As an application of this workflow, a use case demonstrating the assessment of CEACAM6 as a potential blood transcriptional biomarker has been developed [52]. This approach combines literature profiling with analysis of public transcriptome datasets to evaluate a gene's relevance across various diseases and conditions, illustrating the practical application of the skills taught in the toolkit.

 doi:10.20944/preprints202410.0933.v1

These toolkits represent a significant step towards democratizing access to advanced data analysis skills. By providing structured, hands-on training experiences, they enable researchers to develop the capabilities necessary for effective data mining and interpretation, even in resource-limited settings.

The development and dissemination of these workshop toolkits align with the CD2K initiative's goal of empowering researchers to leverage public data effectively. As these resources become more widely available, they have the potential to significantly enhance the data analysis capabilities of the broader biomedical research community, particularly in LMICs.

References

1. Chaussabel D, Rinchai D. Using 'collective omics data' for biomedical research training. Immunology. 2018 Sep;155(1):18-23.
2. Toufiq M, Rinchai D, Bettacchioli E, et al. Harnessing large language models (LLMs) for candidate gene prioritization and selection. J Transl Med. 2023 Oct 16;21(1):728.
3. Preechanukul A, Yimthin T, Tandhavanant S, et al. Abundance of ACVR1B transcript is elevated during septic conditions: Perspectives obtained from a hands-on reductionist investigation. Front Immunol. 2023 Mar 20;14:1072732.
4. Huang SSY, Rinchai D, Toufiq M, et al. Transcriptomic profile investigations highlight a putative role for NUDT16 in sepsis. J Cell Mol Med. 2022 Mar;26(5):1714-1721.
5. Huang SSY, Toufiq M, Saraiva LR, et al. Transcriptome and Literature Mining Highlight the Differential Expression of ERLIN1 in Immune Cells during Sepsis. Biology (Basel). 2021 Aug 5;10(8):755.
6. Toufiq M, Roelands J, Alfaki M, et al. Annexin A3 in sepsis: novel perspectives from an exploration of public transcriptome data. Immunology. 2020 Dec;161(4):291-302.
7. Roelands J, Garand M, Hinchcliff E, et al. Long-Chain Acyl-CoA Synthetase 1 Role in Sepsis and Immunity: Perspectives From a Parallel Review of Public Transcriptome Datasets and of the Literature. Front Immunol. 2019 Oct 18;10:2410.
8. Sawyer AJ, Garand M, Chaussabel D, Feng CG. Transcriptomic Profiling Identifies Neutrophil-Specific Upregulation of Cystatin F as a Marker of Acute Inflammation in Humans. Front Immunol. 2021 Apr 1;12:634119.
9. Le Berre L, Chesneau M, Danger R, et al. Connection of BANK1, Tolerance, Regulatory B cells, and Apoptosis: Perspectives of a Reductionist Investigation. Front Immunol. 2021 Mar 18;12:589786.
10. Riyapa D, Rinchai D, Muangsombut V, et al. Transketolase and vitamin B1 influence on ROS-dependent neutrophil extracellular traps (NETs) formation. PLoS One. 2019 Aug 15;14(8):e0221016.
11. Rinchai D, Kewcharoenwong C, Kessler B, et al. Increased abundance of ADAM9 transcripts in the blood is associated with tissue damage. F1000Res. 2015 Apr 9;4:89.
12. Chaussabel D, Baldwin N. Democratizing systems immunology with modular transcriptional repertoire analyses. Nat Rev Immunol. 2014 Apr;14(4):271-80.
13. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013 Jan;41(Database issue):D991-5.
14. Speake C, Presnell S, Domico K, et al. An interactive web application for the dissemination of human systems immunology data. J Transl Med. 2015 Jun 19;13:196.
15. Bettacchioli E, Chiche L, Chaussabel D, et al. An interactive web application for exploring systemic lupus erythematosus blood transcriptomic diversity. Database (Oxford). 2024 May 28;2024:baae045.
16. Rinchai D, Brummaier T, A Marr A, et al. A data browsing application for accessing gene and module-level blood transcriptome profiles of healthy pregnant women from high- and low-resource settings. Database (Oxford). 2024 Apr 2;2024:baae021.
17. Bougarn S, Boughorbel S, Chaussabel D, Marr N. A curated transcriptome dataset collection to investigate inborn errors of immunity. F1000Res. 2019 Feb 15;8:188.
18. Huang SSY, Al Ali F, Boughorbel S, et al. A curated collection of transcriptome datasets to investigate the molecular mechanisms of immunoglobulin E-mediated atopic diseases. Database (Oxford). 2019 Jan 1;2019:baz066.
19. Bougarn S, Boughorbel S, Chaussabel D, Marr N. A curated transcriptome dataset collection to investigate the blood transcriptional response to viral respiratory tract infection and vaccination. F1000Res. 2019 Mar 13;8:284.
20. Roelands J, Decock J, Boughorbel S, et al. A collection of annotated and harmonized human breast cancer transcriptome datasets, including immunologic classification. F1000Res. 2017 Mar 20;6:296.
21. Mackeh R, Boughorbel S, Chaussabel D, Kino T. A curated transcriptomic dataset collection relevant to embryonic development associated with in vitro fertilization in healthy individuals and patients with polycystic ovary syndrome. F1000Res. 2017 Feb 23;6:181.

 doi:10.20944/preprints202410.0933.v1

22. Rahman M, Boughorbel S, Presnell S, et al. A curated transcriptome dataset collection to investigate the functional programming of human hematopoietic cells in early life. F1000Res. 2016 Mar 30;5:414.
23. Marr AK, Boughorbel S, Presnell S, et al. A curated transcriptome dataset collection to investigate the development and differentiation of the human placenta and its associated pathologies. F1000Res. 2016 Mar 9;5:305.
24. Rinchai D, Boughorbel S, Presnell S, et al. A curated compendium of monocyte transcriptome datasets of relevance to human monocyte immunobiology research. F1000Res. 2016 Apr 25;5:291.
25. Blazkova J, Boughorbel S, Presnell S, et al. A curated transcriptome dataset collection to investigate the immunobiology of HIV infection. F1000Res. 2016 Mar 11;5:327.
26. Toufiq M, Huang SSY, Boughorbel S, et al. SysInflam HuDB, a Web Resource for Mining Human Blood Cells Transcriptomic Data Associated with Systemic Inflammatory Responses to Sepsis. J Immunol. 2021 Nov 1;207(9):2195-2202.
27. Altman MC, Rinchai D, Baldwin N, et al. Development of a fixed module repertoire for the analysis and interpretation of blood transcriptome data. Nat Commun. 2021 Jul 19;12(1):4385.
28. Rinchai D, Roelands J, Toufiq M, et al. BloodGen3Module: blood transcriptional module repertoire analysis and visualization using R. Bioinformatics. 2021 Aug 25;37(16):2382-2389.
29. Bettacchioli E, Chiche L, Chaussabel D, et al. An interactive web application for exploring systemic lupus erythematosus blood transcriptomic diversity. Database (Oxford). 2024 May 28;2024:baae045.
30. Rinchai D, Brummaier T, A Marr A, et al. A data browsing application for accessing gene and module-level blood transcriptome profiles of healthy pregnant women from high- and low-resource settings. Database (Oxford). 2024 Apr 2;2024:baae021.
31. Rawat A, Rinchai D, Toufiq M, et al. A Neutrophil-Driven Inflammatory Signature Characterizes the Blood Transcriptome Fingerprint of Psoriasis. Front Immunol. 2020 Nov 24;11:587946.
32. Rinchai D, Altman MC, Konza O, et al. Definition of erythroid cell-positive blood transcriptome phenotypes associated with severe respiratory syncytial virus infection. Clin Transl Med. 2020 Dec;10(8):e244.
33. Brummaier T, Rinchai D, Toufiq M, et al. Design of a targeted blood transcriptional panel for monitoring immunological changes accompanying pregnancy. Front Immunol. 2024 Jan 30;15:1319949.
34. Rinchai D, Syed Ahamed Kabeer B, Toufiq M, et al. A modular framework for the development of targeted Covid-19 blood transcript profiling panels. J Transl Med. 2020 Jul 31;18(1):291.
35. Toufiq M, Rinchai D, Bettacchioli E, et al. Harnessing large language models (LLMs) for candidate gene prioritization and selection. J Transl Med. 2023 Oct 16;21(1):728.
36. Chaussabel D, Pulendran B. A vision and a prescription for big data-enabled medicine. Nat Immunol. 2015 May;16(5):435-9.
37. Garmire LX, Gliske S, Nguyen QC, et al. The training of next generation data scientists in biomedicine. Pac Symp Biocomput. 2017;22:640-645.
38. Byrd JB, Greene AC, Prasad DV, Jiang X, Greene CS. Responsible, practical genomic data sharing that accelerates research. Nat Rev Genet. 2020 Oct;21(10):615-629.
39. Bourne PE, Bonazzi V, Dunn M, et al. The NIH Big Data to Knowledge (BD2K) initiative. J Am Med Inform Assoc. 2015 Nov;22(6):1114-20.
40. Hancock JM, Zvelebil M, Hollich V, et al. ELIXIR-UK role in bioinformatics training at the national level and across ELIXIR. F1000Res. 2016 Jul 27;5:ELIXIR-952.
41. Mulder NJ, Adebiyi E, Alami R, et al. H3ABioNet, a sustainable pan-African bioinformatics network for human heredity and health in Africa. Genome Res. 2016 Feb;26(2):271-7.
42. Welch L, Lewitter F, Schwartz R, et al. Bioinformatics Curriculum Guidelines: Toward a Definition of Core Competencies. PLoS Comput Biol. 2014 Mar; 10(3): e1003496.
43. Shaikh AR, Butte AJ, Schully SD, et al. Collaborative Biomedicine in the Age of Big Data: The case of cancer. J Med Internet Res. 2014 Apr; 16(4): e101.
44. Payne PRO. Biomedical informatics meets data science: current state and future directions for interaction. *JAMIA Open*, Volume 1, Issue 2, October 2018, Pages 136–141,
45. Chaussabel D, Baldwin N. Democratizing systems immunology with modular transcriptional repertoire analyses. Nat Rev Immunol. 2014 Apr;14(4):271-80.
46. Bourne PE. Ten simple rules for making good oral presentations. PLoS Comput Biol. 2007 Apr;3(4):e77.
47. Attwood TK, Blackford S, Brazas MD, et al. A global perspective on evolving bioinformatics and data science training needs. Brief Bioinform. 2019 Mar 25;20(2):398-404.
48. Strasser BJ, Edwards PM. Big data is the answer ... but what is the question? Osiris. 2017;32:328-345.
49. Mulder N, Schwartz R, Brazas MD, et al. The development and application of bioinformatics core competencies to improve bioinformatics training and education. PLoS Comput Biol. 2018 Feb 1;14(2):e1005772.
50. Al Ali F, Marr AK, Tatari-Calderone Z, et al. Organizing training workshops on gene literature retrieval, profiling, and visualization for early career researchers. F1000Res. 2023 May 11;10:275.

 doi:10.20944/preprints202410.0933.v1

51. Rinchai D, Chaussabel D. A training curriculum for retrieving, structuring, and aggregating information derived from the biomedical literature and large-scale data repositories [version 1; peer review: 2 approved with reservations]. F1000Research. 2022;11:994.
52. Rinchai D, Chaussabel D. Assessing the potential relevance of CEACAM6 as a blood transcriptional biomarker [version 2; peer review: 1 approved, 1 approved with reservations]. F1000Research. 2024;11:1294.