The International Health Cohorts Consortium (IHCC) advances population health research and genomic discovery

John J. Connolly, Scott Sundseth, Grant M. Wood, Chisom Nwaneri, Geoffrey Ginsburg,
Philip Awadalla, Michele Ramsay, Thomas Keane, Peter Goodhand,
Adam S. Butterworth & Hakon Hakonarson

Check for updates

Established in 2018 to push beyond the constraints of individual health and population cohorts, the IHCC is a community of cohorts advancing global science and health. We summarize the collective resources of 69 member cohorts, representing over 34 million people.

Background

The International Health Cohorts Consortium (IHCC) was established in 2018 at the request of the leaders of the Heads of International Research Organizations (HIROs) and through a collaboration between the Global Genomic Medicine Collaborative (G2MC) and the Global Alliance for Genomics and Health (GA4GH). It is a global initiative aimed at closing gaps in genomic databases to enhance representation of different ancestral groups, promoting collaboration across academic and industry partners, and supporting cutting-edge research in areas that impact global health[1]. The mission of the IHCC is to forge cohort connections that revolutionize population health science by providing sustainable data infrastructure, cultivating a collaborative research environment, and promoting policies and best practices that foster connectivity, interoperability, and reciprocity.

IHCC membership criteria prioritize large population cohorts, capability for longitudinal health follow-up, broad participant selection, biological sample collection, and a commitment to data-sharing. It also recognizes the value of smaller cohorts representing underrepresented or unique groups.

Projects supported by the IHCC are focused on the biological, environmental, and social determinants of health and disease. Its core objectives are to (1) accelerate research, (2) harmonize data, (3) educate researchers, (4) enhance public health, and (5) foster innovation. Collaborators can utilize IHCC resources, including the public IHCC Data Atlas (https://atlas.ihccglobal.org/), to drive distinct areas of research in collaboration with other cohorts, and/or to access samples and data from populations of interest.

In this Comment, we summarize the breadth of IHCC members' resources, invite prospective partners to join us in addressing global challenges in health research, and propose a federated template for constructive collaboration.

A large number of IHCC resources are available

Participants from member cohorts span a broad range of ages, ethnicities, and geographic locations, with 35% of locations (N = 24) self-identifying as a low- and middle-income country (LMIC), per World Bank criteria[2].

A members' survey was disseminated from November 2021 to March 2022 and again from November 2022 to March 2023. In total, it was completed by 69 of the 89 cohort members. Respondents reflected the diversity of the IHCC as a whole and were from Africa (N = 7), Asia (N = 19), Australia (N = 2), North America (N = 19), South America (N = 3), and Europe (N = 19). Of the 69 member cohorts who responded, 45 (65%) hold genomic data on some/all of their cohort. Forty-one sites had collected genotype data from ~12,834,036 research participants. Collectively, the cohorts represent ~34 million unique research participants with available data/samples (Fig. 1).

Several million biospecimens are available for immediate research use, subject to project-specific informed consent and ethical oversight, and country-specific legislation. Cohorts with ongoing recruitment continuously add to these collections (Fig. 2).

Relevant phenotype data are largely broad-based (i.e., not phenotype-specific). Demographic data informative of social determinants of health are widely available. Additionally, detailed clinical outcomes address chronic disease, mental health, and lifestyle. Ten (14%) IHCC cohorts (including two LMICs) reported linkage to participants' electronic medical records (EMR).

IHCC encourages data sharing and collaboration

The IHCC supports open and timely access to research findings, adopting the Framework for Responsible Sharing developed by the Global Alliance for Genomics and Health (GA4GH)[3]. This assumes that data donors (or their legal representatives) have provided consent for data use and sharing, and that ethics oversight consistent with local, national, and relevant international laws has been applied, as well as culturally appropriate ethics standards and best practices for governing future data use. The intention and willingness to collaborate with other members and beyond is a prerequisite to membership and fundamental to the IHCC research enterprise.

IHCC facilitates LMIC/LRS representation and support

Since its founding, a core component of the IHCC's mission has been to enhance inclusion of cohorts with diversity from LMICs/low-resource settings (LRS) and to contribute to closing the existing ancestral and minority population gaps in databases. LRSs are characterized by significant constraints in both financial and human resources, encompassing clinical and non-clinical personnel. These contexts are further defined by an underdeveloped organizational and physical infrastructure. In addition to supporting LMIC/LRS collaborations, the IHCC explicitly supports diversity across age, ethnicity, sex, rural/urban, socioeconomic status, both in terms of constituent cohorts and investigator development.
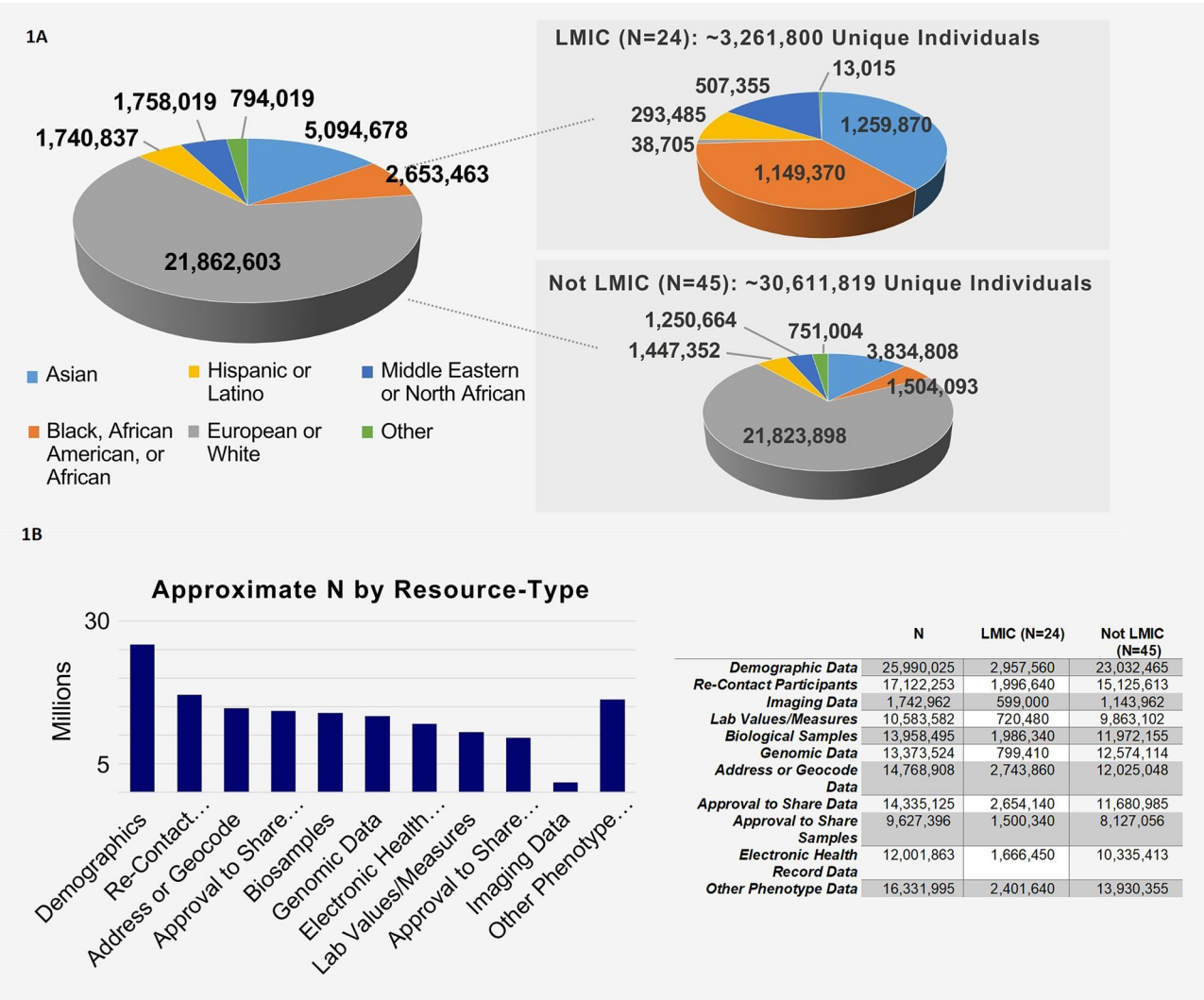
Fig. 1 | Broad breakdown of underlying cohorts as reported by 69 unique IHCC member cohorts. A Breakdown of ~33,872,619 unique participants across 69 reporting sites. Participants are categorized by sub-population for All, LMIC, and non-LMIC member cohorts. **B** Breakdown of resources/datatypes for unique participants as reported by member cohorts.

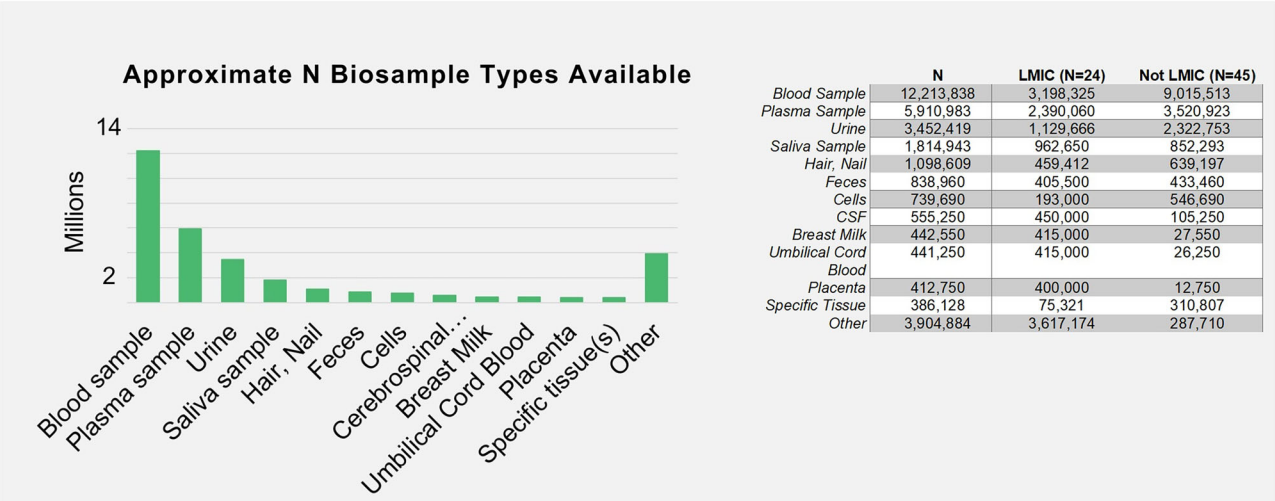## The Global Cohort Atlas allows robust and simple data sharing

The IHCC Global Cohort Atlas serves as a centralized resource for discovering available data across cohorts. It enables data harmonization and sharing across different platforms. Participation in the Atlas is a requirement for any projects funded by the IHCC.

In 2020, the Atlas was launched as the first open global directory of large-scale human cohorts, with cohort information gathered from surveys and data dictionaries. It enables the findability of cohort data by providing users with a single entry point to cross-query cohort data dictionaries to discover cohorts, phenotypes, or variables of interest. Searchable dimensions include participant disease status, data use policy, sample collection parameters, genotypes, and demographic/health variables from cohort data dictionaries. The Atlas encourages data interoperability by providing

harmonized descriptions of cohorts. All of the Atlas's cohort public metadata (cohort descriptors, data dictionaries, and variables) are openly available (https://github.com/IHCC-cohorts).

Cohorts are classified into three levels of detail based on the cohort variables collected and harmonized. High-level cohort fields are collected upon joining IHCC (Level 1), structured cohort descriptors are collected through the IHCC annual cohort surveys (Level 2), and complete semantic harmonization of the cohort data dictionary variables is carried out in collaboration with cohort data managers (Level 3). This multi-level approach has enabled us to lower the bar for inclusion in the Atlas (Level 1), balancing the curation overhead required to reach the most comprehensive Level 3.

The Cohort Atlas has grown to 89 cohorts from 43 countries with 12 harmonized cohorts (Level 3) through collaborations with projects such as CINECA and the Davos Alzheimer's Collaborative (DAC).

Approximate N Biosample Types Available

	N	LMIC (N=24)	Not LMIC (N=45)
Blood Sample	12,213,838	3,198,325	9,015,513
Plasma Sample	5,910,983	2,390,060	3,520,923
Urine	3,452,419	1,129,666	2,322,753
Saliva Sample	1,814,943	962,650	852,293
Hair, Nail	1,098,609	459,412	639,197
Feces	838,960	405,500	433,460
Cells	739,690	193,000	546,690
CSF	555,250	450,000	105,250
Breast Milk	442,550	415,000	27,550
Umbilical Cord Blood	441,250	415,000	26,250
Placenta	412,750	400,000	12,750
Specific Tissue	386,128	75,321	310,807
Other	3,904,884	3,617,174	287,710

Fig. 2 | Breakdown of approximate number and types of biosamples available as reported by member sites. Across the 69 responding cohorts, millions of biospecimens are available for immediate research use, subject to project-specific informed consent and ethical oversight, and country-specific legislation. Cohorts with ongoing recruitment continuously add to these sample collections, ensuring that both new and existing studies can access fresh or archived samples as needed.

Table 1 | The IHCC-DAC model for program development

Stage	Description
I	*Design and implement a pilot program* that will serve as a proof-of-concept for feasibility and delineate requisite resources/strategies that would support a large-scale program. An online survey assessed a) data availability among IHCC cohorts, and b) feasibility to generate new data.
II	*Establish a high-priority cohort based on polygenic risk scores (PRSs)*: Individuals at highest risk are targeted for deep phenotyping—beginning with what exists readily and correlates with dementia risk. Examples include hypertension, heart failure, CAD/PAD, CVAs (stroke/MI), type 2 diabetes, and obesity—all of which can be collected from ICD and billing codes and can be validated through re-contact (where consent allows) for deeper molecular phenotyping.
III	*Identify high-risk individuals with pre-dementia or early dementia symptoms* in the age range of 45–70 years who can be invited to undergo additional workup/testing, including imaging and biomarker studies.
IV	*Prioritize pursuit of cohorts with patient re-contact and IRB-approved broad-based data sharing plans (or such feasibility) across all races and ethnicities*: Racial and ethnic minority populations, underserved populations, and populations who experience poorer medical outcomes have been underrepresented in genomic research to date. This is widely recognized as detrimental to efforts to catalog and interpret genomic variants and use them to improve clinical care. Addressing these health disparities is considered fundamental to the goals of IHCC and is similarly a key focus of the proposed project with DAC.

An important endpoint of the initiative was to identify high-priority participants. These participants can be subsequently pursued for generation and analyses of deep-phenotype data, with follow-up (as consent allows) to conduct network-wide analyses of deeper data.

Starting a collaboration

Collaborations typically start with a feasibility assessment through the Atlas and outreach to cohort leaders for project alignment.

The Alzheimer's pilot aimed to integrate existing cohort data worldwide to form a global Alzheimer's disease (AD) resource. By uniting diverse cohorts with varying degrees of genomic and phenotypic data, the program targeted early detection of disease onset and progression. The approach to project planning, initiation, and development constitutes a working model for future IHCC programs with a milestone-driven approach to program development (Table 1).

In the planning stages, over 70 experts defined specific recommendations for the 'Must have' minimum set of components and approaches, as well as suggestions for desirable "Nice to have" components. A full mapping and assessment of existing cohorts of subjects willing and able to contribute, already enrolled participants, was undertaken. Twenty IHCC cohorts comprising several million individuals were identified. Each completed a resource survey aligned with the scientific plan, which allowed the IHCC to assess readiness, gaps, and support needs. In addition, the group cataloged additional cohorts/participants that offer congruent capabilities, as well as specialized sub-cohorts focused on narrow aspects such as brain autopsy or novel biomarker strategies. In the pilot phase, the IHCC pursued cohort leads from cohorts to contribute data for the assessment of AD polygenic risk in diverse populations in comparison with European ancestry cohorts, including the UK Biobank[4].

We envision that this model is adaptable to any number of phenotypes/projects, whereby global data on diverse and heterogeneous populations will yield new discoveries.

Existing pilots and publications

Our existing pilots give an example of the diversity of areas in which our research can be used. Several cohorts are engaged in using PRS to assess genetic susceptibility to various diseases, including cardiovascular diseases, diabetes, and dementia. These pilot studies are crucial for understanding PRS application in underrepresented populations[5]. Similarly, several cohorts are focused on mental health disorders, including studies to identify genetic and environmental contributors to conditions such as depression,

schizophrenia, and anxiety, particularly in non-European populations[6–9]. The IHCC continues to support studies pioneering the use of metabolomics in population health. This research is crucial for understanding how metabolic pathways contribute to chronic diseases such as diabetes and cardiovascular conditions in diverse populations[10,11]. Another example of IHCC-supported health initiatives is in studying the impact of opioids—known carcinogens—on cancer burden. Member cohorts have formed a global collaboration to harmonize data on opioid and cancer risk[12,13]. Similarly, IHCC Cohorts are collaborating on the study of early-life adiposity linked to cancer risk in diverse genetic ancestries[14–16]. Finally, cohorts from Africa have developed a resource for coronavirus host genomics studies. This multi-collaborator strategic partnership was designed to provide harmonized demographic, clinical, and genetic information specific to Black South Africans with COVID-19[17].

Future research directions

We aim to continue our support for the pilot program and to expand to new areas in the coming year. Our subgroups remain active and continue to address gaps in global health research. Future efforts will focus on improving the understanding of mental health disorders across LMIC populations, where the burden of mental health conditions is rising but remains under-researched. We also plan to enhance members' metabolomics capabilities, particularly by developing standards for cross-cohort data integration. New initiatives include unique opportunities to further assess the relationship between climate and health. With over 70% of IHCC cohorts collecting environmental data, we are well-powered to lead in this increasingly important domain. Other areas for IHCC expansion include cancer genomics, infectious disease research, and pharmacogenomics. These fields represent new frontiers where diverse global data will be important for identifying genetic and environmental contributors to health.

Conclusions

The IHCC is a unique resource and transformative global initiative aimed at uniting large-scale, longitudinal population cohort studies to address some of the most pressing challenges in population health. The consortium's commitment to diversity, inclusion, and representation addresses existing gaps in genomic databases and health disparities. Through initiatives such as the Global Cohort Atlas, standardized data-sharing frameworks, and federated analysis models, the IHCC has enabled collaborations that transcend geographical and technological barriers while respecting ethical and regulatory considerations.

IHCC's scientific strategy has catalyzed novel research, including the application of PRS, studies on mental health and metabolomics, and the development of global frameworks for Alzheimer's disease research. As it continues to grow, the IHCC remains focused on advancing its mission to support sustainable, inclusive, and high-impact research.

The IHCC invites researchers, policymakers, and industry partners to join its efforts in building a collaborative ecosystem that leverages population cohort data to drive innovation and equity in health outcomes worldwide. Researchers interested in joining can contact ihccinfo@ihccglobal.org.

Data availability

All of the underlying cohort public metadata, including cohort descriptors, data dictionaries, and variables, are openly available. Collaborators can utilize IHCC resources, including the public IHCC Data Atlas at: https://atlas.ihccglobal.org/. All of the Atlas's cohort public metadata (cohort descriptors, data dictionaries, and variables) are openly available: https://github.com/IHCC-cohorts.

John J. Connolly [iD][1,2] [✉], **Scott Sundseth**[3], **Grant M. Wood**[3], **Chisom Nwaneri**[3], **Geoffrey Ginsburg**[4], **Philip Awadalla**[5], **Michele Ramsay** [iD][6], **Thomas Keane** [iD][7], **Peter Goodhand**[8], **Adam S. Butterworth** [iD][9,10,11,12,13] & **Hakon Hakonarson** [iD][1,2,14,15]

[1]Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia (CHOP), Philadelphia, PA, USA. [2]Department of Pediatrics, the Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA. [3]Global Genomic Medicine Collaborative, Durham, NC, USA. [4]All of Us Research Program, National Institutes of Health, Bethesda, MD, USA. [5]Nuffield Department of Population Health, University of Oxford, Big Data Institute, Oxford, UK. [6]Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa. [7]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton Cambridge, UK. [8]Global Alliance for Genomics and Health; Ontario Institute for Cancer Research, Toronto, ON, Canada. [9]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [10]Victor Phillip Dahdaleh Heart and Lung Research Institute, University of Cambridge, Cambridge, UK. [11]British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, UK. [12]National Institute for Health and Care Research Blood and Transplant Research Unit in Donor Health and Behaviour, University of Cambridge, Cambridge, UK. [13]Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK. [14]Divisions of Human Genetics and Pulmonary Medicine, Children´s Hospital of Philadelphia, Philadelphia, USA. [15]Faculty of Medicine, University of Iceland, Reykjavik, Iceland. [✉]e-mail: connollyj1@chop.edu

References

1. Manolio, T. A., Goodhand, P. & Ginsburg, G. The International Hundred Thousand Plus Cohort Consortium: integrating large-scale cohorts to address global scientific challenges. *Lancet Digit. Health* **2**, e567–e568 (2020).
2. Organization WH. *The Global Health Observatory* (World Health Organization, 2024).
3. Knoppers, B. M. Framework for responsible sharing of genomic and health-related data. *Hugo J.* **8**, 3 (2014).
4. Sleiman, P. M. et al. Trans-ethnic genomic informed risk assessment for Alzheimer's disease: an International Hundred K+ Cohorts Consortium study. *Alzheimers Dement.* **19**, 5765–5772 (2023).
5. Qu, H. Q. et al. Trans-ethnic polygenic risk scores for body mass index: an international hundred K+ cohorts consortium study. *Clin. Transl. Med.* **13**, e1291 (2023).
6. Brunoni, A. R. et al. Prevalence and risk factors of psychiatric symptoms and diagnoses before and during the COVID-19 pandemic: findings from the ELSA-Brasil COVID-19 mental health cohort. *Psychol. Med.* **53**, 446–457 (2023).
7. Choi, K. W. et al. Effects of social support on depression risk during the COVID-19 pandemic: what support types and for whom? Preprint at medRxiv https://doi.org/10.1101/2022.05.15.22274976 (2022).
8. Fatori, D. et al. Trajectories of common mental disorders symptoms before and during the COVID-19 pandemic: findings from the ELSA-Brasil COVID-19 Mental Health Cohort. *Soc. Psychiatry Psychiatr. Epidemiol.* **57**, 2445–2455 (2022).
9. Lee, Y. H. et al. Association of everyday discrimination with depressive symptoms and suicidal ideation during the COVID-19 pandemic in the All of Us research program. *JAMA Psychiatry* **79**, 898–906 (2022).
10. Qu, H. Q. et al. Metabolomic profiling for dyslipidemia in pediatric patients with sickle cell disease, on behalf of the IHCC consortium. *Metabolomics* **18**, 101 (2022).
11. Qu, H. Q. et al. Metabolomic profiling of samples from pediatric patients with asthma unveils deficient nutrients in African Americans. *iScience* **25**, 104650 (2022).
12. Sheikh, M., Brennan, P., Mariosa, D. & Robbins, H. A. Opioid medications: an emerging cancer risk factor? *Br. J. Anaesth.* **130**, e401–e403 (2023).

13. Alcala, K. et al. Incident cancers attributable to using opium and smoking cigarettes in the Golestan cohort study. *EClinicalMedicine* **64**, 102229 (2023).
14. Papadimitriou, N. et al. Separating the effects of early and later life adiposity on colorectal cancer risk: a Mendelian randomization study. *BMC Med.* **21**, 5 (2023).
15. Papadimitriou, N. et al. Body mass index at birth and early life and colorectal cancer: a two-sample Mendelian randomization analysis in European and East Asian genetic similarity populations. *Pediatr. Obes.* **20**, e13186 (2025).
16. Mukhtar, M. et al. The Associations of selenoprotein genetic variants with the risks of colorectal adenoma and colorectal cancer: case-control studies in Irish and Czech populations. *Nutrients* **14**, 2718 (2022).
17. May, A. K. et al. Coronavirus Host Genomics Study: South Africa (COVIGen-SA). *Glob. Health Epidemiol. Genom.* **2022**, 7405349 (2022).

Author contributions

John J Connolly: Conception, writing, and data curation. Scott Sundseth: Project administration, writing, review, and editing. Grant M. Wood: Project administration. Chisom Nwaneri: Project administration. Geoffrey Ginsburg: Conception, review, and editing. Philip Awadalla: Review and editing. Michele Ramsay: Review and editing. Thomas Keane: Review and editing. Peter Goodhand: Conception, review, and editing. Adam S. Butterworth: Conception, review, and editing. Hakon Hakonarson: Conception, review, and editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43856-025-01026-y.

Correspondence and requests for materials should be addressed to John J. Connolly.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.