GLOBAL BIODATA COALITION

Global Biodata Coalition

Highlights 2024

This publication was produced by :
The Global Biodata Coalition

Cover illustration and brochure design :
Spencer Phillips

For more information about the Global Biodata Coalition please contact :
info@globalbiodata.org

Printer :
Newprint

Table of Contents

Describing the Global Biodata Infrastructure: updating the global inventory

Introduction

One of GBC's activities is to understand the size and scope of the global infrastructure of biodata resources: how many resources are there, where are they, and which resources are of core importance for researchers? In 2022 we began addressing the first two questions by developing an open and reproducible machine learning-enabled pipeline to identify biodata resources described in the scientific literature from 2011–2021. We reported in the 2022 GBC Highlights brochure that the inventory identified 3012 distinct resources distributed around the world and published a description of the methodology and results in 2023. In 2024 we published two papers: one describing the **inventory and its results** and a second detailing how we implemented a reproducible workflow to enable periodic repetition of the inventory, allowing us to understand how the infrastructure changes over time.

Reproducibility and FAIRness

Scientific reproducibility is a persistent challenge, particularly in computational research. Many studies lack the necessary documentation, code accessibility, and standardised methodologies required for others to replicate findings. To address this, the GBC implemented a rigorously designed workflow, ensuring that every step of the inventory process—from data collection to model training and validation—was transparent and replicable. Our approach to implementing this reproducible workflow is described in a **2024 publication**. In addition to ensuring reproducibility, a core principle of the inventory project was adherence to the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles. To achieve this:

- All code and data were openly shared on **GitHub** and **Zenodo**, enabling other

researchers to verify and build upon the work.

- Data interoperability was prioritised, with structured formats ensuring ease of integration with existing databases.
- Reproducibility was built-in through comprehensive documentation, automated workflows, and external validation of methods.

Re-running the inventory workflow

In 2024, we re-ran the inventory pipeline to include biodata resources described in the scientific literature from 2011–2023: including two additional years from the original work. This work successfully demonstrated the reproducibility of the workflow, and also found 3773 biodata resources, or 761 more than we found in 2022—a 24% increase in just two years.

These results indicate that the biodata infrastructure is dynamic, and growing rapidly–an indicator, we think, of the growth of life sciences research and of the importance of generating, archiving, and analysing those data for the research endeavour. The inventory workflow and dataset is open and available on the GBC's GitHub repository for all to use. GBC is currently working to further analyse these results to understand where these biodata resources are located, who funds them, and how they, or even if, they are cited in the scientific literature.

The Broader Impact

By making the inventory workflow fully reproducible, the GBC has set a precedent for future research in computational biodata analysis. This initiative serves as a model for other scientific fields, demonstrating how machine learning and open science practices can enhance the transparency and reliability of research. Furthermore, the updated inventory provides valuable insights for funding bodies, policymakers, and the broader scientific community, supporting efforts to sustain and strengthen global biodata infrastructure.

Photo credit : Jeff Dowling

Foreword

2024 has been a year of significant progress for the **Global Biodata Coalition** as we continue to deepen our understanding of the global biodata infrastructure and reinforce the essential role of helping to sustain biodata resources that play such an important part in advancing science, health, and innovation. Our efforts to support long-term sustainability have taken shape through key initiatives that highlight both the breadth and the impact of biodata resources.

A major achievement this year was the successful repeat of the Global Inventory of Biodata Resources, allowing us to refine and expand our insight into the evolving landscape of global biodata infrastructure. In parallel, we published six **impact studies** demonstrating how biodata underpins key developments in health, the environment, and the economy, reinforcing the critical role of these resources in global research and policy-making.

Looking ahead to 2025, we anticipate the launch of the third Global Core Biodata Resource (GCBR) selection round, further strengthening the recognition and coordination of key biodata resources. In the summer our Sustainability Lab will bring together experts from our community, as well as from the social and behavioural sciences, to consider potential models for the sustainability of global biodata resources. Additionally, in October, the GBC's funders and biodata resource managers will participate in a joint meeting of the GBC Board and the GCBR Forum, fostering greater collaboration on the tackling of sustainability challenges.

As we move forward, we invite all stakeholders to engage with us in shaping a robust, sustainable future for the global biodata ecosystem.

Dr Guy Cochrane
Executive Director

Examining the impact of global biodata resources across the sectors

The GBC commissioned six comprehensive studies to evaluate the multifaceted impact of biodata resources across a range of sectors, with a particular focus on the Global Core Biodata Resources that the GBC considers to be of fundamental importance to the wider biological and life sciences community and the long term preservation of biological data. These impact studies underscore the indispensable role of biodata in advancing scientific research, healthcare, environmental conservation, and economic development.

1. Improving Scientific Practices

Biodata resources are pivotal in standardising methodologies, ensuring data quality, and promoting reproducibility in scientific research. By providing accessible databases and tools, these resources facilitate consistent data annotation and sharing, thereby enhancing the reliability and efficiency of scientific investigations.

View Here

2. Economic Activity and Industrial Innovation

The integration of biodata resources into industrial applications has spurred economic growth by enabling the development of novel products, technologies, and services. Industries such as biotechnology, pharmaceuticals, and agriculture

Impact Study

Children's Health

Impact Study

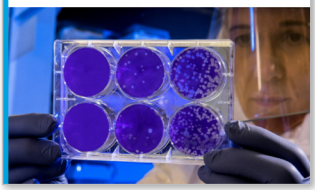
Economic Activity & Industrial Applications

Impact Study

Environmental & Biodiversity Preservation

Impact Study

Public Health & Therapeutics



Impact Study

Rare Diseases & Genetic Diagnostics

Impact Study

Improving Scientific Practices

leverage these resources to innovate and maintain competitiveness in the global market.

View here

3. Children's Health

Advances in digital infrastructure and biodata resources have markedly improved health outcomes for children. Notably, the five-year survival rate for childhood cancer has increased significantly over the past five decades, reflecting the critical role of biodata in understanding disease processes, early diagnosis, personalised treatment plans, and ongoing health monitoring.

View here

4. Environmental and Biodiversity Preservation

Biodata resources contribute significantly to environmental conservation efforts by providing comprehensive data on biodiversity and ecosystem changes. These resources inform policy decisions and conservation strategies, addressing challenges such as habitat alteration and oceanic changes, thereby aiding in the preservation of global biodiversity.

View here

5. Public Health and Therapeutics

In the realm of public health, biodata resources have been instrumental in accelerating research and enhancing responses to health crises. By facilitating rapid data sharing and analysis, these resources support the development of effective therapeutics and inform public health strategies to manage and mitigate disease outbreaks.

View here

6. Rare Diseases and Enhancing Genetic Diagnostics

A study involving over 740 families with rare diseases demonstrated that whole genome sequencing, supported by biodata resources, was essential in identifying disease-causing variants that other methods missed. This highlights the critical role of biodata in advancing genetic

diagnostics and providing insights into rare diseases, ultimately guiding more effective treatments.

View here

Collectively, these studies affirm the transformative impact of biodata resources across multiple domains. The GBC remains committed to supporting and expanding these resources to foster innovation, improve health outcomes, and promote sustainable development worldwide.

The GBC acknowledges the work of Research Consulting in the production of these impact stories.

GBC's *Impact Study* series.

The GBCR Forum in 2024

In 2024, the **Global Core Biodata Resource (GCBR) Forum** continued to serve as a vital platform for discussion and collaboration among the world's leading biodata resources. Meeting quarterly, the Forum brought together **managers of the 52 Global Core Biodata Resources** to exchange insights and discuss challenges to supporting their resources. These meetings facilitated critical discussions on emerging opportunities and risks for biodata resources, particularly in relation to funding sustainability, technological advancements, and the evolving landscape of data usage.

A major discussion point in 2024 was the rapid development of **Artificial Intelligence (AI) and Large Language Models (LLMs)** and the opportunities and challenges for biodata resources. While AI offers potential efficiency gains in data curation and accessibility to be realised in the longer term future, Forum members highlighted key challenges, including the need for human validation to ensure accuracy, the complexity of integrating AI into existing workflows, and concerns over AI-generated data potentially compromising trust in curated knowledgebases. Several GCBRs shared their experiences in applying AI to areas such as literature mining, helpdesk automation, software development and data annotation. The Forum also acknowledged that while AI can enhance resource usability, its implementation requires additional time, funding, and oversight to maintain high data quality standards. This is an issue on which GBC will continue to engage biodata resources and funders, including through a workshop at Biocuration 2025 that GBC is co-organising with Forum members.

Another key discussion area centered on **improving the design of funding calls** to better meet the needs of biodata resources. Members noted that current funding mechanisms often prioritise innovation over ongoing service provision, leading to challenges in securing stable support for core infrastructure.

The Forum explored ways to address this imbalance, including decoupling operational funding from innovation grants, developing clearer guidance for grant reviewers assessing biodata resource proposals, and fostering cross-funder collaboration to align expectations across different funding bodies. Additionally, members discussed how funders could engage commercial users, particularly in the pharmaceutical sector, to contribute equitably to biodata resource sustainability while maintaining open data principles.

A third major discussion topic considered **tracking the usage and impact of Global Core Biodata Resources**. Forum members examined current approaches to usage metrics, including text and data mining of citations, as well as the integration of Research Organization Registry (ROR) identifiers to track institutional engagement. Members agreed on the importance of developing standardised usage tracking approaches, both to help funders understand the critical role of biodata resources and to ensure resources can effectively demonstrate their impact on scientific research.

Looking ahead, the GCBR Forum will continue to serve as a crucial mechanism for international cooperation and strategic dialogue around biodata resource sustainability. In 2025, the Forum will continue, through it's Working Group to define elements of sustainability as well as to hold its first joint meeting with the GBC's Board of Funders. By addressing these challenges collectively, the Forum reinforces its commitment to securing a sustainable future for the global life sciences community in tandem with global life science funding bodies.

2024 in Numbers

306,340

times GCBR's have been cited

936,953

times inventory resources have been cited

3,763

resources in our global inventory, hosted in 58 different countries, between 2011 - 2023

Inventory funded by

2,629

different funding agencies or streams from

80

different countries

72

new resources coming online in 2024

Highlights of the Year

JANUARY 2024
Community consultation for GBC White paper

MARCH 2024
Meeting with Advisor to the Department of Biotechnology at the Indian Biological Data Centre in Faridibad

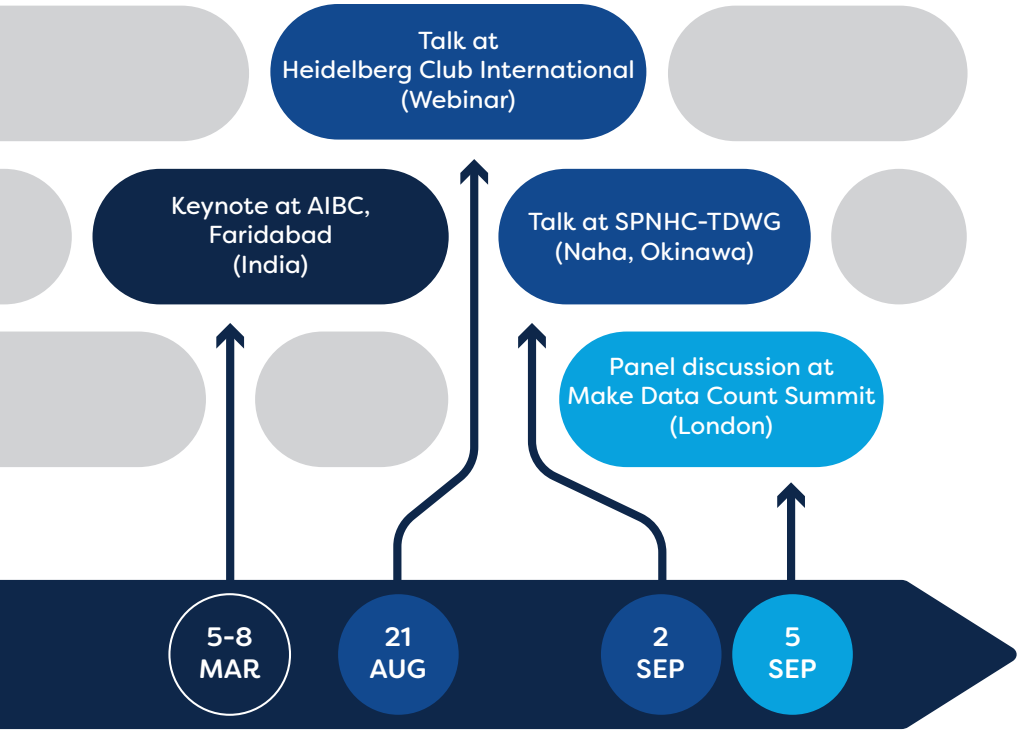
JULY 2024
Re-run of Inventory process, identifying 761 new biodata resources

NOVEMBER 2024
GBC Open Letter Campaign launches

GBC Around the World

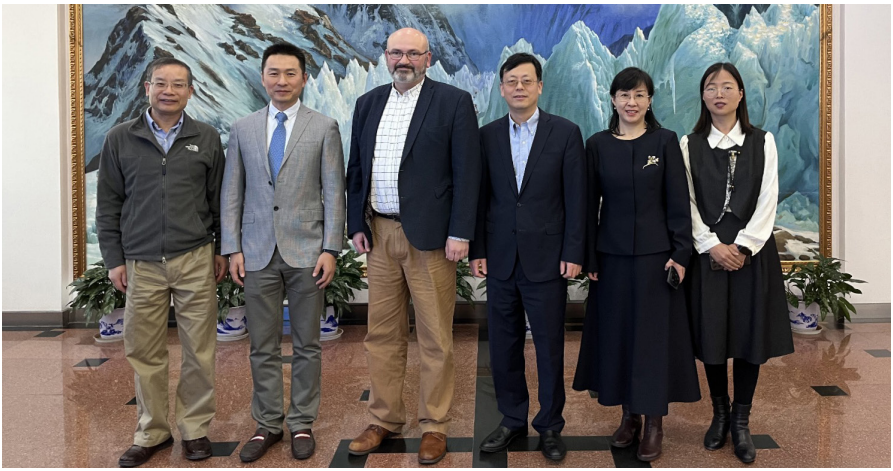
The GBC's **outreach programme**, spearheaded by Executive Director, Guy Cochrane and Program Manager, Chuck Cook, saw engagement with representatives of global life science funding organisations through a number of bilateral meetings.

Additionally, we continued to engage stakeholders through presentations at relevant scientific conferences and meetings and the convening of panel discussions around the world, both in person and virtually.



GBC Executive Director, Guy Cochrane gives keynote address at 17th Annual International Biocuration Conference (AIBC-2024), Faridabad, India.

Delegates at the 17th Annual International Biocuration Conference.

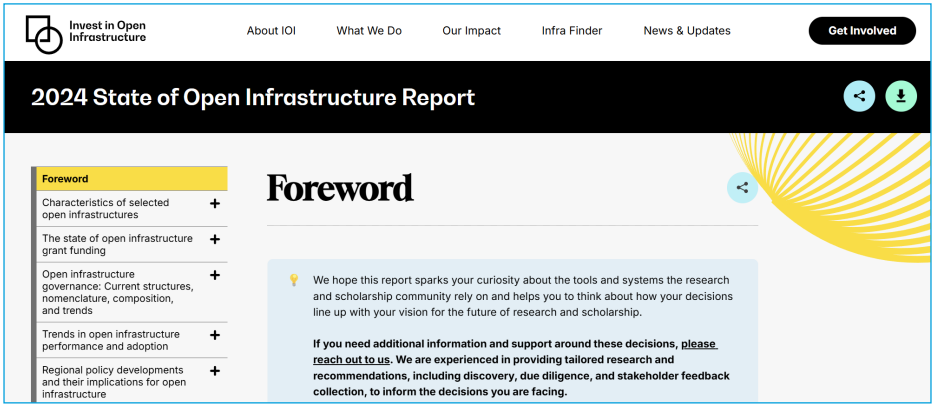


GBC Executive Director, Guy Cochrane meets officials from the Chinese Academy of Sciences (CAS), Beijing, China.

GBC in the News

The GBC plays a vital role in advancing the sustainability of global biodata resources, and its inclusion in open publications underscores the growing recognition of its work. Being featured in reports, studies and news articles enhances awareness of GBC's mission, reinforcing the importance of coordinated international efforts in biodata management.

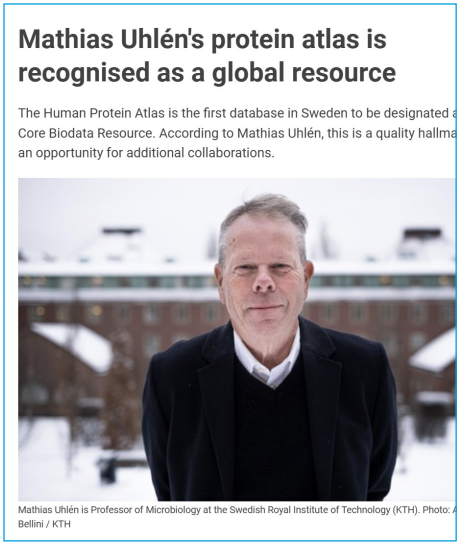
In April, IDW, the German science news service reported on the Leibniz Institute's DSMZ being the host of two newly selected Global Core Biodata Resources: **SILVA** and **LPSN**. SILVA provides quality-controlled ribosomal RNA gene sequences for microorganism identification, while LPSN is the authoritative source on prokaryotic nomenclature.
In June 2024, the State of Open Infrastructure report, published by the Invest in Open Infrastructure organisation, highlighted GBC's contributions to open infrastructure funding and governance. The report recognised GBC's role in fostering transparency, collaboration, and



Below: The State of Open Infrastructure report, published by the Invest in Open Infrastructure Organisation.

sustainable policies—aligning with broader global trends in open science and infrastructure sustainability.

Subsequent to the conclusion of the its second Global Core Biodata Resource selection round, the GBC was referenced in **Life Science Sweden** in recognition of its role in designating the **Human Protein Atlas (HPA)** as a Global Core Biodata Resource (GCBR). This acknowledgment highlights HPA's significance as a high-impact, globally used biological database. The article underscores the importance of GBC's selection process, which identifies key biodata resources essential to life science research. According to HPA's founder, Mathias Uhlén, the designation serves as a hallmark of quality and fosters further global collaborations.



Mathias Uhlén's protein atlas is recognised as a global resource

The Human Protein Atlas is the first database in Sweden to be designated a Core Biodata Resource. According to Mathias Uhlén, this is a quality hallma an opportunity for additional collaborations.

Mathias Uhlén is Professor of Microbiology at the Swedish Royal Institute of Technology (KTH). Photo: A Bellini / KTH

Left: Life Science Sweden recognises the Human Protein Atlas's (HPA) designation as a Global Core Biodata Resource (GCBR).

Forward Look

2025 & Beyond

Looking ahead to 2025, the Global Biodata Coalition (GBC) remains committed to advancing dialogue and action among its stakeholder community of funders, resource managers and others about the sustainability and coordination of the world's most critical biodata resources.

In the summer we will hold the first-of-a-kind meeting entitled **"Sustainability Lab"** where we will foster new partnerships and explore innovative funding models by bringing together diverse specialists from within and outside the life science data community to consider innovative and potentially disruptive new biodata funding models. We will take forward the learning from the Sustainability Lab and other recommendations derived from the Sustainability Working Group of the GCBR Forum to better define biodata sustainability models, securing the infrastructure for the future.

An important focus in 2025 will be the further strengthening of the existing body of 52 Global Core Biodata Resources (GCBRs).

We will open a call for expressions of interest for our third GCBR selection round which will provide an opportunity for biodata resources that meet the GCBR criteria, but were not able to engage in the two earlier processes, to participate. Later in the year, the **GBC's Board of Funders** will hold its first joint in-person meeting with the GCBR Forum which will provide an opportunity for a deep dive into developing solutions and shaping the policies that will underpin the future of open biological data.

Another priority for the coming year is to continue to raise awareness of the GBC globally. To that end, in April we will lead a workshop at the International Society of Biocuration (ISB) conference in Kansas City, Missouri in which we will exchange experiences, and encourage discussion of AI in curation of biodata resources. We will also be present at the biennial joint ISMB ECCB (Intelligent Systems for Molecular Biology / European Conference on Computational Biology) meeting that will take place in July in Liverpool and subsequently the Genomic Standards Consortium (GSC) annual meeting in Cambridge.

GBC Members in 2024

Members

- US NIH-NHGRI
- US NSF
- Wellcome
- UKRI - Biotechnology and Biological Sciences Research Council
- Chan Zuckerberg Initiative
- EMBL
- Forskningsrådet
- GenomeCanada
- The African Academy of Sciences
- SERI, Switzerland

Observers

- European Commission
    - European Research Council
- Inserm (France)
- Korean National Institutes of Health
- South African Medical Research Council

In-kind Support

We are grateful for in-kind support for the GBC from :

- The Human Frontier Science Program Organization

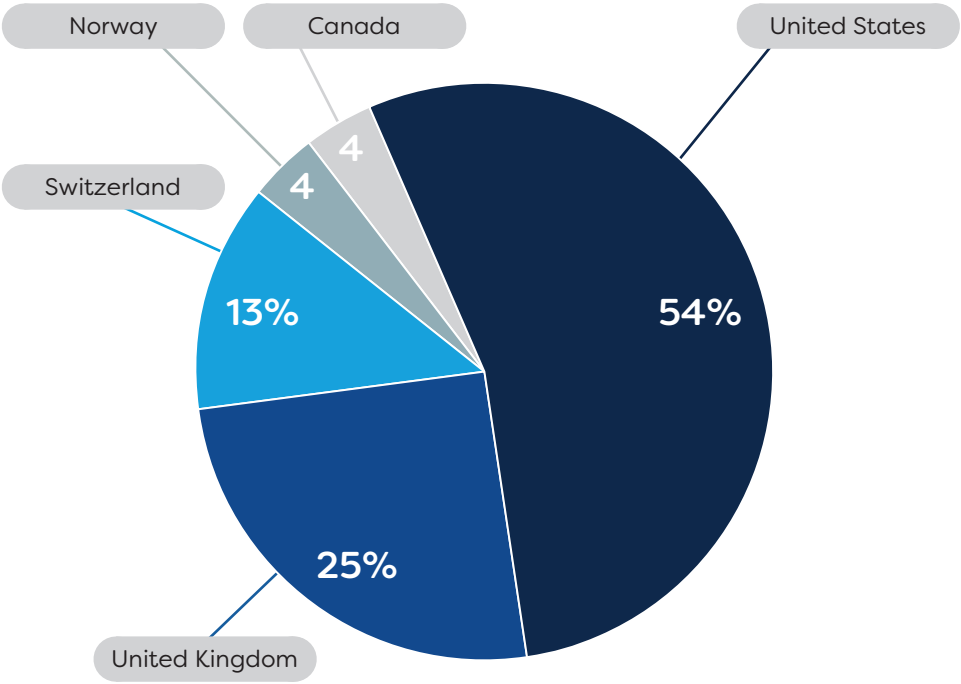
Financial Figures

Funding sources / Operating expenditure

Funding

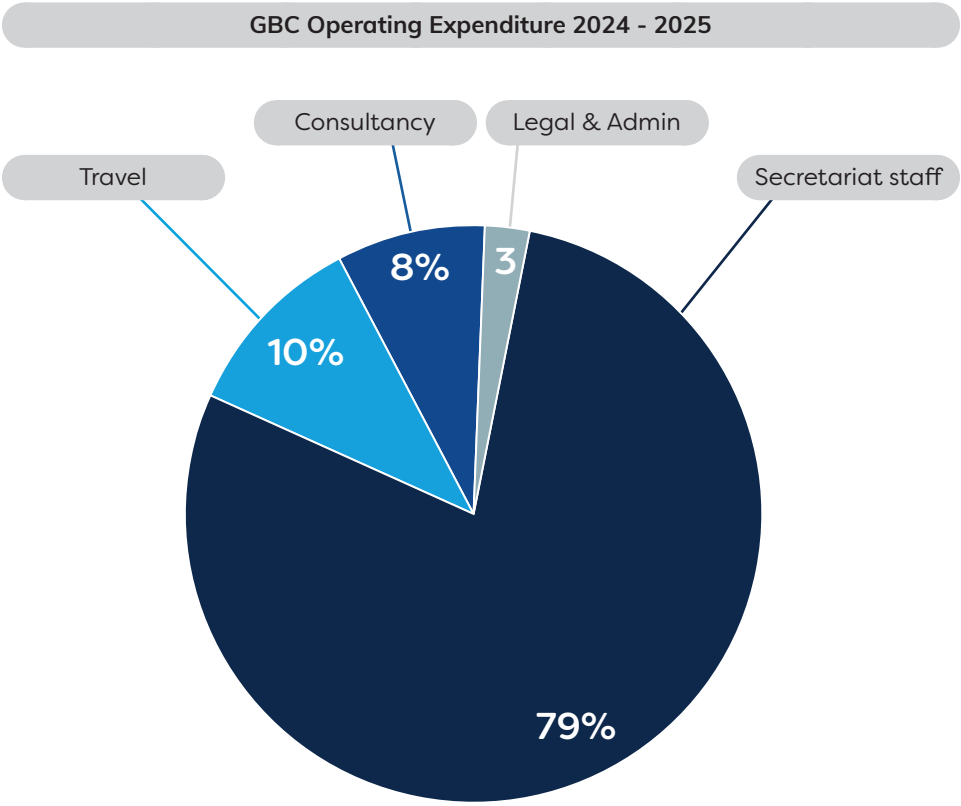
The GBC Secretariat is funded by contributions from its Members, who are public and charitable research funding organisations. In 2024, 58% of our funding was from North America and 42% from Europe.

GBC Annual Contributions by Region



Operating Expenditure

GBC's fiscal year runs from April to March. In the 2024-2025 fiscal year we estimate, using expenditure from the first nine months of the year, that GBC's total expenditure will be approximately €550,000. Of this, 79% supports the secretariat, 8% is for external consultants, 10% for travel, and 3% for administrative costs.



Global Biodata Coalition

www.globalbiodata.org

info@globalbiodata.org

@globalbiodata

www.linkedin.com/company/global-biodata-coalition

A digital version of this publication is available at :

www.globalbiodata.org/news-and-resources/