Global Genomic
Medicine Collaborative

# 2nd International Cohorts Summit

**April 23–24, 2019**

**Grand Hotel Reykjavík**
**Reykjavík, Iceland**

International Cohorts Summit

# Table of Contents

# Summit Objectives

The International Cohorts Summit was conceived by the **Heads of International Research Organizations (HIROs)** chaired by Jeremy Farrar of the Wellcome Trust. An inaugural summit was held in March 2018 in Durham, North Carolina, USA. It was organized by the Global Genomic Medicine Collaborative (G2MC, www.g2mc.org) with the goal of enabling leaders of large-scale longitudinal cohorts worldwide to share best practices, discuss data sharing, explore standards, discuss common challenges, and explore the potential for a larg(er) collaborative sequencing strategy. From that Summit, the International Hundred K+ Cohorts Consortium (IHCC) was formed, with teams assembled to address the actions identified in this Summit.

The second International Cohorts Summit, held in Reykjavik, Iceland, was attended by 117 people with representation from 67 cohorts from 29 countries. The summit continued the discussion to develop the scientific agenda, drive the research and collaboration between cohorts, and present team activities and progress on goals.

The second International Cohorts Summit was held in Reykjavik Iceland with 117 in attendance, with representation from 67 cohorts from 29 countries

The primary objectives of this meeting were to

- Identify scientifically meritorious cross-cohort research projects and international collaborators willing to organize and participate in them
- Develop an IHCC scientific agenda to bring forward to funding bodies

We thank you for joining us at this pivotal event and for contributing to our shared goal of enhanced international collaboration across research projects.

# Programme Committee

Philip Awadalla, Ph.D., *Ontario Institute for Cancer Research, Canada*
Adam Butterworth, Ph.D., *University of Cambridge School of Clinical Medicine, UK*
Zhengming Chen, D.Phil., *China Programmes, China*
Robert Eiss, *Fogarty International Center, USA*
Paul Elliott, Professor, M.B.B.S., Ph.D., F.Med.Sci., *Imperial College London, UK*
Geoff Ginsburg, M.D., Ph.D., *Duke University and G2MC, USA*
Peter Goodhand, *Global Alliance for Genomics & Health, Canada*
Hakon Hakonarson, M.D., Ph.D., *The Joseph Stokes Jr. Research Institute of Children's Hospital of Philadelphia, USA*
Norihiro Kato, M.D., Ph.D., *National Center for Global Health and Medicine, Japan*
Thomas Keane, Ph.D., *European Bioinformatics Institute, UK*
Gun Peggy Knudsen, Ph.D., *Norwegian Institute of Public Health, Norway*
Rongling Li, M.D., Ph.D., M.P.H., *National Human Genome Research Institute, USA*
Laura Lyman Rodriguez, Ph.D., *National Human Genome Research Institute, NIH, USA*
Teri Manolio, M.D., Ph.D., *National Human Genome Research Institute and G2MC, USA*
Teji Rakhra-Burris, *Duke University and G2MC, USA*
Gad Rennert, M.D., Ph.D., *Carmel Medical Center, Israel*

Global Genomic Medicine Collaborative

International Cohorts Summit

# International Cohorts Summit

Hosted by the Global Genomic Medicine Collaborative (G2MC)

***Grand Hotel Reykjavík***
***Reykjavík, Iceland***
***April 23—24, 2019***

**Meeting Objectives**

- Identify scientifically meritorious cross-cohort research projects and identify international collaborators willing to organize and participate in them
- Develop an IHCC scientific agenda to bring forward to funding bodies

## Day 1, Tuesday, April 23, 2019

| | | |
|---|---|---|
| 7:30 - 8:30 am | **REGISTRATION** | Gullteigur Reception |
| 7:30 - 8:30 am | **HOTEL BREAKFAST** | Grand Brasserie |

### Session 1 – Introduction and Background
**Chair: Geoff Ginsburg**

| | |
|---|---|
| 8:30 - 8:35 am | **Welcome and Introductions**<br>*Geoffrey Ginsburg, Peter Goodhand, and Teri Manolio*<br>Duke University; Global Alliance for Genomics & Health; National Human Genome Research Institute<br>USA and Canada |
| 8:35 - 8:45 am | **Welcoming Remarks**<br>*Guðni Thorlacius Jóhannesson*<br>President of Iceland |
| 8:45 - 8:50 am | **Keynote Introduction**<br>*Hákon Hákonarson*<br>Children's Hospital of Philadelphia<br>USA |

| | | |
|---|---|---|
| 8:50 - 9:15 am | **Keynote**: **Using Big Data as a Role Model to Improve Population Health in Iceland**<br>*Birgir Jakobsson*<br>Medical Advisor to the Minister of Health<br>Iceland | |
| 9:15 - 9:35 am | **Cohorts: A Collective Vision**<br>*Mary De Silva*<br>Wellcome Trust<br>UK | |
| 9:35 - 9:45 am | **Innovations in Cohort Study Methods in *All of Us***<br>*Kelly Gebo*<br>National Institutes of Health *All of Us* Research Program<br>USA | |
| 9:45 - 10:05 am | **Overview of Current Progress**<br>*Geoff Ginsburg*<br>Duke University<br>USA | |
| 10:05 - 10:30 am | **BREAK** | |

## Session 2 – Opportunities for Partnership
### Chair: Peter Goodhand

| | | |
|---|---|---|
| 10:30 - 10:45 am | **Cohorts in the EU and Beyond**<br>*Barbara Kerstiëns*<br>Directorate-General for Research and Innovation<br>European Commission<br>Belgium | |
| 10:45 - 11:00 am | **Moving from Genome Discoveries to Disease Mechanisms**<br>*Nancy Cox*<br>Vanderbilt University Medical Center<br>USA | |
| 11:00 - 11:15 am | **The European Project SYNCHROS**<br>*Josep Marie Haro*<br>Parc Sanitari Sant Joan de Déu<br>Spain | |
| 11:15 am - 12:15 pm | **LUNCH** | Grand Brasserie |

## Session 3 – Data Standards and Infrastructure
### Chairs: Philip Awadalla & Thomas Keane

| | |
|---|---|
| 12:15 - 12:30 pm | **Vision and Challenges to Achieve Cohort Interoperability**<br>*Philip Awadalla and Thomas Keane*<br>Canadian Partnership for Tomorrow Project, Global Alliance for Genomics & Health; European Bioinformatics Institute<br>Canada; UK |
| 12:30 - 12:45 pm | **The Genomic Aetiology of Osteoarthritis**<br>*Ele Zeggini*<br>Institute of Translational Genomics<br>Germany |
| 12:45 - 1:00 pm | **Data Challenges in Sharing Across Cohorts**<br>*Nicola Mulder*<br>University of Cape Town<br>South Africa |
| 1:00 - 2:20 pm | **RFI Presentations** |
| 1:00 - 1:20 pm | **Genetic and Non-Genetic Risk Factors for Uncommon Cardiometabolic Conditions**<br>*Adam Butterworth*<br>University of Cambridge<br>UK |
| 1:20 - 1:40 pm | **A Pilot Study of Data Harmonization and Rare Variant Detection among International Cohorts**<br>*Rongling Li*<br>National Human Genome Research Institute<br>USA |
| 1:40 - 2:00 pm | **High-Throughput Metabolomic Biomarker Measures across IHCC Cohorts**<br>*Hákon Hákonarson*<br>Children's Hospital of Philadelphia<br>USA |
| 2:00 - 2:20 pm | **Toward a Federated Data Ecosystem**<br>*Anthony Philippakis*<br>Broad Institute<br>USA |
| 2:20 - 2:35 pm | **BREAK** |

## Session 4 – Scientific Strategy and Cohorts Enhancement
### Chairs: Adam Butterworth, Hákon Hákonarson, & Gad Rennert

2:35 - 2:50 pm **Opportunities and Deliverables in Scientific Strategy**
*Adam Butterworth, Hákon Hákonarson, and Gad Rennert*
University of Cambridge; Children's Hospital of Philadelphia; Carmel Medical Center/Technion–Israel Institute of Technology
UK; USA; Israel

2:50 - 3:05 pm **FinnGen & Global Biobank Analysis**
*Mark Daly*
Institute for Molecular Medicine Finland
Finland

3:05 - 3:20 pm **Body Mass Index and Mortality: Using Big Data for Common Conditions**
*Emanuele Di Angelantonio*
University of Cambridge
UK

3:20 - 4:40 pm **RFI Presentations**

3:20 - 3:40 pm **A Global Understanding of the Role of Proteomic and Metabolomic Profiles in Health and Disease**
*Heather Eliassen and Fran Grodstein*
Harvard T.H. Chan School of Public Health, Harvard University; Brigham and Women's Hospital
USA

3:40 - 4:00 pm **Pharmacogenomic Analysis Across IHCC Diversity Cohorts for Assessment of Drug Response**
*Kenny Nguyen and Hákon Hákonarson*
Children's Hospital of Philadelphia
USA

4:00 - 4:20 pm **International 100k+ Consortium (IHCC) Drug Development Resource**
*Aroon Hingorani*
University College London
UK

4:20 - 4:30 pm **The Human Knockout Project**
*Daniel MacArthur*
Harvard University/Broad Institute/Massachusetts General Hospital
USA

4:30 - 4:40 pm **Identification of Loss of Function (LOF) Variants in Health and Disease**
*Hákon Hákonarson*
Children's Hospital of Philadelphia
USA

4:40 - 4:55 pm **BREAK**

## Session 5 – Policy and Bio-Data Sharing
### Chairs: Gun Peggy Knudsen & Laura Lyman Rodriguez

| | |
|---|---|
| 4:55 - 5:05 pm | **Recap of Analyses to Date and Planning for Next Steps**<br>*Gun Peggy Knudsen and Laura Lyman Rodriguez*<br>Norwegian Institute of Public Health; National Human Genome Research Institute<br>Norway; USA |
| 5:05 - 5:20 pm | **Large Scale Openly Accessible Blood-based Epidemiological Studies in India**<br>*Prabhat Jha*<br>University of Toronto<br>Canada |
| 5:20 - 5:40 pm | **About European Research Infrastructures & Implementing the GDPR: Towards a Code of Conduct for Health Research**<br>*Michaela Mayrhofer*<br>BBMRI-ERIC<br>Austria |
| 5:40 - 7:00 pm | **RFI Presentations** |
| 5:40 - 6:00 pm | **Application of Polygenic Risk Scores (PRS) for Improved Health Outcomes in Alzheimer Disease, Cardiovascular Disease and Across Multiple Neuropsychiatric-Phenotypes**<br>*Patrick Sleiman and Hákon Hákonarson*<br>Children's Hospital of Philadelphia<br>USA |
| 6:00 - 6:20 pm | **A Precision Medicine Approach to Multi-morbidity in Cardio-Metabolic Disease and Dementia**<br>*Sarah Bauermeister on behalf of John Gallacher*<br>University of Oxford<br>UK |
| 6:20 - 6:40 pm | **Addressing the Ethnicity Gap in Human Genetics Research**<br>*Adam Butterworth*<br>University of Cambridge<br>UK |
| 6:40 - 7:00 pm | **Atopic Diseases, Early Life Exposure Risk Factors, and Genetics in the Young (ALERGY) Study**<br>*Elizabeth Jensen*<br>Wake Forest School of Medicine<br>USA |
| 7:00 - 7:10 pm | **Closing Remarks**<br>*Teri Manolio*<br>National Human Genome Research Institute<br>USA |

| | | |
|---|---|---|
| 7:10 pm | **ADJOURN** | |
| 7:30 pm | Depart for dinner [meet in lobby of Grand Hotel] | |
| 8:00 pm | Group Dinner [Út í Bláinn, Perlan] | |

# Day 2, Wednesday, April 24, 2019

| | | |
|---|---|---|
| 7:00 - 8:00 am | **REGISTRATION** | Gullteigur Reception |
| 7:00 - 8:00 am | **HOTEL BREAKFAST** | Grand Brasserie |
| 8:00 - 8:05 am | **Keynote Introduction**<br>*Peter Goodhand*<br>Global Alliance for Genomics & Health<br>Canada | |
| 8:05 - 8:35 am | **Keynote: Big Data in Health Care and Research, Opportunities and Challenges**<br>*Ewan Birney*<br>EMBL-European Bioinformatics Institute<br>UK | |
| 8:35 - 8:45 am | **Charge to Breakout Groups**<br>*Geoff Ginsburg*<br>Duke University<br>USA | |

## Session 6 – Break Out Sessions (2 hours)

**8:45-10:45 am**

| | | |
|---|---|---|
| Team A | **Data Standards and Infrastructure**<br>*Philip Awadalla and Thomas Keane*<br>Canadian Partnership for Tomorrow Project, Global Alliance for Genomics & Health; European Bioinformatics Institute<br>Canada; UK | Hvammur |
| Team B | **Scientific Strategy and Cohorts Enhancement**<br>*Adam Butterworth, Hákon Hákonarson, and Gad Rennert*<br>University of Cambridge; Children's Hospital of Philadelphia; Carmel Medical Center/Technion–Israel Institute of Technology<br>UK; USA; Israel | Gullteigur |
| Team C | ***Policy and Bio-Data Sharing***<br>*Gun Peggy Knudsen and Laura Lyman Rodriguez*<br>Norwegian Institute of Public Health; National Human Genome Research Institute<br>Norway; USA | Galleri |

10:45 - 11:15 am **BREAK**

## Session 7 - Team Leads Synthesis and LUNCH

11:15 am - 12:45 pm **Team Leads Synthesis Breakout Session** Hvammur

11:15 am - 12:45 pm **LUNCH** Grand Brasserie

## Session 8 – Break Out Reports and Discussion
### Moderator: Teri Manolio

12:45 - 1:15 pm **Data Standards and Infrastructure**
*Philip Awadalla and Thomas Keane*
Canadian Partnership for Tomorrow Project, Global Alliance for Genomics & Health; European Bioinformatics Institute
Canada; UK

1:15 - 1:45 pm **Scientific Strategy and Cohorts Enhancement**
*Adam Butterworth, Hákon Hákonarson, and Gad Rennert*
University of Cambridge; Children's Hospital of Philadelphia; Carmel Medical Center/Technion–Israel Institute of Technology
UK; USA; Israel

1:45 - 2:15 pm **Policy and Bio-Data Sharing**
*Gun Peggy Knudsen and Laura Lyman Rodriguez*
Norwegian Institute of Public Health; National Human Genome Research Institute
Norway; USA

2:15 - 2:45 pm **Cross-Team Synergies**
*Team Leads*

2:45 – 3:00 pm **BREAK**

## Session 9 – Summary and Action Planning
### Chairs: Geoff Ginsburg, Peter Goodhand, & Teri Manolio

3:00 – 4:00 pm **Summary, Consensus and Next Steps**
*Geoff Ginsburg, Peter Goodhand, and Teri Manolio*
Duke University; Global Alliance for Genomics & Health; National Human Genome Research Institute
USA and Canada

4:00 pm **ADJOURN**

4:00 pm **NETWORKING BREAK** Gullteigur Reception

## CINECA (Common Infrastructure for National Cohorts in Europe, Canada, and Africa) Project Satellite Session

**Chair: Nicky Mulder**

| | | |
|---|---|---|
| 4:30 - 6:00 pm | **CINECA Stakeholder Engagement Session**<br>The project aims to develop tools for federated data analysis across cohorts. In addition to the 13 participating cohorts, they wish to engage other cohorts to determine the bottlenecks in cross-cohort data analysis and to gather information on training needs in this area.<br><br>**Open to all Cohorts Summit Attendees* | Gullteigur |

# Summit Presenters & Session Chairs

**Gudni Th. Johannesson**

President of Iceland

**Gudni Th. Johannesson** is the sixth president of Iceland. He was elected in 2016 and became Iceland's youngest president at 48 years old.

President Johannesson received his bachelor's degree in political science at Warwick University, his master of studies degree in history from Oxford University, and his doctorate in history from Queen Mary, University of London. Before taking office, Johannesson was a professor of history at the University of Iceland. He has written several books on modern Icelandic history, a book about spying in Iceland, and a book about the 2008 banking collapse. He has also written dozens of scholarly and newspaper articles. In 2017, he was awarded an honorary degree by Queen Mary University.

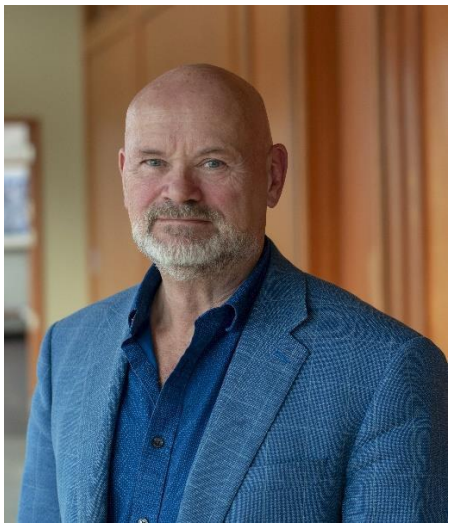
## IHCC Co-Chairs

**Geoff Ginsburg**

Director
Center for Applied Genomics & Precision Medicine
Duke University Medical Center
USA

**Geoff Ginsburg** is the founding director for the Center for Applied Genomics & Precision Medicine at the Duke University Medical Center and for MEDx, a partnership between the Schools of Medicine and Engineering to spark and translate innovation. His research addresses the challenges for translating genomic information into medical practice and the integration of precision medicine into healthcare. In 2017 he received Duke's Translational Research Mentorship Award.

He is a member of the Advisory Council to the Director of NIH and is co-chair of the National Academies Roundtable on Genomic and Precision Health and is founder and president of the Global Genomic Medicine Collaborative, a not-for-profit organization aimed at creating international partnerships to advance the implementation of precision medicine. He has recently served as a member of the Board of External Experts for the National Heart, Lung, and Blood Institute, the advisory council for the National Center for Advancing Translational Sciences, the chair of the review for Genome Canada's Large Scale Applied Research Competition in Genomics and Precision Medicine, and the World Economic Forum's Global Agenda Council on the Future of the Health Sector.



**Peter Goodhand**

Chief Executive Officer
Global Alliance for Genomics and Health (GA4GH)
Canada

**Peter Goodhand** is the Chief Executive Officer of the Global Alliance for Genomics and Health (GA4GH), as well as a leader in the global health sector as a senior executive and board member. Additionally, he has fifteen years of experience as a patient advocate, caregiver, and navigator throughout his family's battle with a rare cancer.

Goodhand is currently a member of the Occupational Cancer Research Centre Steering Committee, Co-Chair of the Medical and Scientific Advisory Board of Global Genes, Co-Chair of the International 100K+ Cohorts Consortium (IHCC), and a member of the Global Genomic Medicine Collaboration (G2MC) Steering Committee.

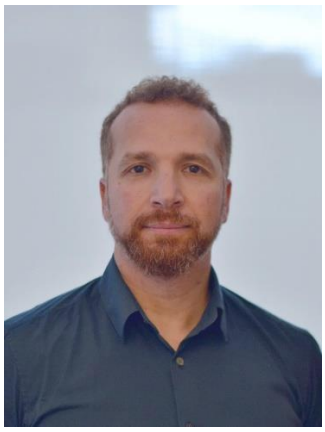
**Teri Manolio**

Director, Division of Genomic Medicine
National Human Genome Research Institute (NHGRI)
USA

**Teri Manolio, M.D., Ph.D.,** is a physician and epidemiologist at NIH and Walter Reed National Military Medical Center. She joined NHGRI in 2005 to lead efforts in applying genomic technologies to population research. She has authored over 270 research reports and has research interests in genome-wide association studies of complex diseases, ethnic differences in disease risk, and incorporating genomic findings into clinical care.

As a physician and epidemiologist, Dr. Manolio has a deep interest in discovering genetic changes associated with diseases by conducting biomedical research on large groups of people. As the director of the Division of Genomic Medicine, Dr. Manolio leads efforts to support research translating those discoveries into diagnoses, preventive measures, treatments and prognoses of health conditions.

## Presenters & Session Chairs



### Philip Awadalla

National Scientific Director
Canadian Partnership for Tomorrow Project (CPTP)
Canada

**Philip Awadalla, Ph.D.**, is the national scientific director for the Canadian Partnership for Tomorrow Project, director of Computational Biology and the executive scientific director of Ontario Health Study at the Ontario Institute for Cancer Research, and professor of Population and Medical Genomics at the University of Toronto. He is director of the Genome Canada, Canadian Data Integration Centre. He obtained his doctorate in population and statistical genetics from the University of Edinburgh and was awarded NSERC, Killam, and Wellcome Trust Fellowships to pursue his postdoctoral work before taking faculty positions at North Carolina and the University of Montreal. He was the scientific director of CARTaGENE, and was part of the analysis groups of the 1000 Genomes Project and PCAWG. Major projects include genomics of aging, hematological diseases and cancers. Other projects include estimating mutation and recombination rates and model-based approaches to identify genetic and environmental control points for infectious diseases in Africa.

### Sarah Bauermeister

Senior Researcher-Psychometric Analyst, European Prevention of Alzheimer's Dementia (EPAD)
Senior Data Manager – Cognitive Neuropsychologist Dementias Platform UK (DPUK)
UK

**Sarah Bauermeister, Ph.D.**, is a senior researcher working across multiple projects as a psychometric analyst for the European Prevention of Alzheimer's Disease (EPAD) study and a multi-disciplinary analyst for Dementias Platform UK (DPUK) in the research group of Professor John Gallacher, director of DPUK. She is also a senior data manager for DPUK, a data repository platform for over 2 million participants across 45+ population and disease-specific cohorts (https://www.dementiasplatform.uk/). She is the programme lead for the DPUK datathon intiative, lead for the DPUK user group - DPUK Reach, programme lead for the DPUK data curation project and scientific lead for an international DPUK collaborative AD project. Dr. Bauermeister is also a scientific reviewer for DPUK Data Portal research applications.

Her psychometric work includes conducting Item Response Theory (IRT) analyses on mental health, lifestyle and healthcare scale data, investigating item-level information and scale reliability. She also utilizes the Structural Equation Modelling (SEM) to explore comorbidities such as cognition and mental health, child adversity and adult biomedical outcomes and dementia. Her previous research focused on the cognitive predictors of falls and frailty in older adults, and poor mental health as a comorbidity with cognitive decline and dementia.

**Ewan Birney**

Joint Director
The European Bioinformatics Institute
(EMBL-EBI)
UK

**Ewan Birney, Ph.D.,** is director of EMBL-EBI, with Dr. Rolf Apweiler, and runs a small research group. He is also EMBL-EBI's joint head of research, alongside Dr. Nick Goldman.

Dr. Birney completed his Ph.D. at the Wellcome Sanger Institute with Richard Durbin. In 2000, he became head of nucleotide data at EMBL-EBI and in 2012 he took on the role of associate director at the institute. He became director of EMBL-EBI in 2015. Dr. Birney led the analysis of the Human Genome gene set, mouse and chicken genomes and the ENCODE Project, focusing on non-coding elements of the human genome. Ewan's main areas of research include functional genomics, DNA algorithms, statistical methods to analyse genomic information (in particular information associated with individual differences in humans and Medaks fish) and use of images for chromatin structure.

Dr. Birney is a non-executive director of Genomics England and a consultant and advisor to a number of companies, including Oxford Nanopore Technologies, Dovetail Genomics, and GSK. Dr. Birney was elected an EMBO member in 2012, a fellow of the Royal Society in 2014, and a fellow of the Academy of Medical Sciences in 2015.

He has received a number of awards, including the 2003 Francis Crick Award from the Royal Society, the 2005 Overton Prize from the International Society for Computational Biology and the 2005 Benjamin Franklin Award for contributions in Open Source Bioinformatics.

**Adam Butterworth**

Reader in Molecular Epidemiology
University of Cambridge
UK

**Adam Butterworth, Ph.D.,** is Reader in Molecular Epidemiology at the Department of Public Health and Primary Care, University of Cambridge (UK). Following training in genetics and genetic epidemiology in Cambridge and Sheffield, Adam joined the Cardiovascular Epidemiology Unit in 2009. His main interests are discovering genetic risk factors for cardiovascular diseases (eg, Howson et al., *Nat Genet*, 2017; Nelson et al., *Nat Genet*, 2017) and molecular intermediate traits, such as proteins (Sun et al., *Nature*, 2018) and blood cell traits (Astle *et al., Cell*, 2016). These discoveries can then be used to glean insights into the causal role of biological pathways in cardiovascular and other diseases (eg, Burgess et al., *JAMA Cardiol*, 2018). Adam also has an interest in how genetics might be useful in clinical practice or improving health, such as by genotyping blood groups and providing genetic risk information to motivate behaviour change.

**Nancy Cox**

Founding Director
Vanderbilt Genetics Institute
Vanderbilt University
USA

**Nancy J. Cox, Ph.D.,** is a quantitative human geneticist with a long-standing research program in identifying and characterizing the genetic component to common human diseases. She earned a B.S. in biology from the University of Notre Dame in 1978 and a Ph.D. in human genetics from Yale in 1982, and she did post-doctoral research at Washington University (1982-85) and the University of Pennsylvania (1985-87) before joining the University of Chicago where she was a faculty member in the Departments of Medicine and Human Genetics for 28 years. In 2015 she joined Vanderbilt as the founding Director of the Vanderbilt Genetics Institute and the Mary Phillips Edmonds Gray Professor of Genetics and Medicine. Dr. Cox runs a computational lab focused primarily on integrating genome variation with genome function and with the medical phenome using BioVU, the biobank at Vanderbilt University. Current funded research includes studies in neuropsychiatric disorders, Alzheimer's disease, the genetic component to health disparities, diabetes and its complications, and analysis center activities for the Genomic Sequencing Projects.

**Mark Daly**

Director
Institute for Molecular Medicine Finland
(FIMM)
Finland

**Mark Daly, Ph.D.,** was appointed Director of the Institute for Molecular Medicine Finland (FIMM) in February 2018. He retains also his affiliations in Boston at the Harvard Medical School, Massachusetts General Hospital and as an institute member of the Broad Institute and co-director of the Program in Medical and Population Genetics.

Daly received his B.S. in physics from the Massachusetts Institute of Technology and his Ph.D. in human genetics from Leiden University, Netherlands. He received the Curt Stern Award from the American Society of Human Genetics in 2014 and is a member of the National Academy of Medicine.

His research primarily focuses on the development and application of statistical methods for the discovery and interpretation of genetic variation responsible for complex human disease. In addition to foundational work in human genetics methodology, his lab has made major contributions to gene discovery in inflammatory bowel disease, autism and schizophrenia - primarily through catalyzing global collaborative research efforts which he continues to help lead. He is a co-architect of the FinnGen project, a landmark public-private effort to integrate decades of medical registry data with genomic data in 10% of the Finnish population.

Mark Daly has been an author on more than 450 peer-reviewed manuscripts with a total of more than 200,000 citations, has an h-index of 178, and has been listed by Thompson ISI/Science Watch in 2008 and 2010 as one of the top ten authors ranked by number of high-impact papers. Mark was the recipient of the Curt Stern Award from the American Society of Human Genetics in 2014 and was elected to the National Academy of Medicine in 2017.

**Mary De Silva**

Head of Population Health
The Wellcome Trust
UK

**Mary De Silva, M.Sc., Ph.D.,** joined Wellcome in November 2015 as head of population health, where she directs Wellcome's funding of population health research in the UK and in low- and middle-income countries. Dr. De Silva also co-leads an initiative to develop a new priority area for Wellcome in mental health research, which will launch in 2019. In 2016, Dr. De Silva led the development of a new Longitudinal Studies Strategy for Wellcome, which sets out criteria for sustainably funding large-scale resources and improving access to and use of these resources.

Dr. De Silva has an undergraduate degree in biological anthropology from Cambridge and an M.Sc. and Ph.D. in epidemiology from the London School of Hygiene and Tropical Medicine.

She was previously the deputy director of the Centre for Global Mental Health at the London School of Hygiene and Tropical Medicine, where her research interests included the design and evaluation of complex interventions to improve mental health in low- and middle-income countries, implementation research to ensure that these interventions are scalable and sustainable, and policy influence work to encourage the translation of evidence into policy and practice. She co-founded and led the Mental Health Innovation Network, a global network of researchers, practitioners and policy makers that aims to promote the use of evidence based interventions to improve mental health.

**Emanuele Di Angelantonio**

Director, National Institute of Health Research Blood and Transplant Research Unit in Donor Health and Genomics in Cambridge
Deputy Director of the Cardiovascular Epidemiology Unit in the Department of Public Health at the University of Cambridge
UK

**Emanuele Di Angelantonio, M.Sc., Ph.D.,** is the director of the National Institute of Health Research Blood and Transplant Research Unit in Donor Health and Genomics in Cambridge and the deputy director of the Cardiovascular Epidemiology Unit in the Department of Public Health at the University of Cambridge.

Since 2012, he has been the principal investigator in donor health research and honorary consultant for NHS Blood and Transplant, and in 2018 he was appointed as the inaugural NHSBT Professor of Donor Health.

**Heather Eliassen**

Associate Professor
Harvard Medical School
USA

**Heather Eliassen, Sc.D.,** is an associate professor of Medicine and Epidemiology at Harvard Medical School and the Harvard T.H. Chan School of Public Health. Her research focuses on the etiology of breast cancer, examining associations between lifestyle factors, biomarkers of lifestyle and hormones, and breast cancer risk. She investigates how women may alter their lifestyle to reduce breast cancer risk, and much of her work has been conducted within the Nurses' Health Study (NHS) and Nurses' Health Study II (NHSII). She is co-PI of the NHSII, founded in 1989 by Dr. Walter Willett, that includes over 116,000 women. She also is director of the BWH/Harvard Cohort Biorepository, which houses more than 3 million biospecimens from 200,000 cohort participants, and co-lead of the Cancer Epidemiology Program at the Dana Farber/Harvard Cancer Center. Dr. Eliassen is actively involved in teaching and mentoring graduate students at the Harvard T.H. Chan School of Public Health and mentoring postdoctoral fellows and junior faculty members at Harvard and Brigham and Women's Hospital.

**John Gallacher**

Principal Investigator (PI) and Director,
MRC-funded Dementias Platform UK
Professor of Cognitive Health,
University of Oxford
UK

**Professor John Gallacher** is the Principal Investigator (PI) and Director of the MRC-funded Dementias Platform UK as well as Professor of Cognitive Health at the University of Oxford; visiting Professor to Imperial College London and Honorary Professor at the University of Hong Kong. He is also PI for the Caerphilly Prospective Study, developing the study's focus on ageing and dementia. As member of the UK Biobank Steering Group, he leads on cognitive and psychological assessment.

**Kelly Gebo**

Chief Medical and Scientific Officer
National Institutes of Health *All of Us*
Research Program
USA

**Kelly Gebo, M.D., M.P.H.,** is the Chief Medical and Scientific Officer of the *All of Us* Research Program. In this role, she works with diverse stakeholders to lead the Program's scientific agenda and guide protocol revisions and data collection processes. She also provides clinical oversight of the Program in collaboration with the institutional review board (IRB) and the *All of Us* team.

Dr. Gebo has clinical, research, and educational experience in the health care and higher education sectors. She is a professor of medicine at Johns Hopkins University and an expert in HIV health services research and clinical outcomes of persons with HIV. Previously, she served as the co-principal investigator of the HIV Research Network, an 18-year clinical cohort study of high-volume HIV sites caring for over 20,000 persons with HIV across the country. She has also served as an American Council on Education Fellow at the University of Pennsylvania and as the Vice Provost for Education at Johns Hopkins.

Dr. Gebo holds a doctorate in medicine from the Johns Hopkins University School of Medicine and a master's in public health from the Johns Hopkins Bloomberg School of Public Health.

**Francine Grodstein**

Professor
Brigham and Women's Hospital
Harvard Medical School
USA

**Francine Grodstein, Sc.D.,** received her doctorate in epidemiology in 1992 from Harvard School of Public Health. She is a professor of medicine at Brigham and Women's Hospital, Harvard Medical School. Dr. Grodstein is Director of Research for the Nurses' Health Study, one of the largest studies of women's health.

Dr. Grodstein's research activities over the last 25 years have focused on health in aging. She has published more than 200 original manuscripts and book chapters in this area. She is an advisor to the American Nurses' Association's Program "Healthy Nurse Healthy Nation" and an advisor to the director, Health Disparities Subcommittee, Centers for Disease Control and Prevention. In addition, Dr. Grodstein is a member of the Steering Committee for the NIH Women's Health Initiative Memory Study, a member of the Steering Committee for the Massachusetts Alzheimer Disease Research Center, and on the Scientific Review Board for the Alzheimer's Drug Discovery Foundation. Dr. Grodstein has been an advisor for the U.S. Centers for Disease Control and Prevention's Healthy Brain Initiative. Dr. Grodstein is a member of the Editorial Board of *Menopause*, the journal of the North American Menopause Society, as well as the Executive Board of *Maturitas*, the journal of the European Menopause Society.

**Hákon Hákonarson**

Director of the Center for Applied Genomics
Endowed Chair in Genomics Research
Professor of Pediatrics at the University of Pennsylvania
Perelman School of Medicine
USA

**Hákon Hákonarson, M.D., Ph.D.**, is director of the Center for Applied Genomics, Endowed Chair in Genomics Research, and professor of pediatrics at the University of Pennsylvania, Perelman School of Medicine. Dr. Hákonarson received his M.D. and Ph.D. from the University of Iceland's School of Medicine. He leads a major commitment from Children's Hospital of Philadelphia to genomically characterize approximately 100,000 children, an initiative that has gained nationwide attention in the *Wall Street Journal*, *New York Times*, *Time Magazine*, *Nature*, and *Science*. Dr. Hakonarson is a principal investigator within the Kids First program and the TOPMed genomics program funded by the National Institutes of Health (NIH).

Dr. Hákonarson has previously held several senior posts within the biopharmaceutical industry, directing a number of genomics and pharmacogenomics projects as vice president of Clinical Sciences and Development at deCODE genetics, Inc. Dr. Hákonarson has been the principal investigator (PI) on multiple NIH-sponsored grants, and he was a principal investigator on the Neurodevelopmental Genomics: Trajectories of Complex Phenotypes, the largest research project ever supported by the National Institute of Mental Health. Dr. Hákonarson recently completed a clinical biomarker study in ADHD demonstrating strong efficacy and safety of a neuromodulator compound (NFC-1) in children with specific mutations in the glutamate metabotropic (mGluR) receptor family of genes with ADHD and Autism (www.ClinicalTrial.gov; Elia et al, Nat Comm, 2018).

Dr. Hákonarson has published more than 650 scientific papers, including numerous high-impact papers on genomic discoveries and their translations in some of the most prestigious scientific medical journals, including *Nature*, *Nature Medicine, Nature Genetics, Cell,* and *The New England Journal of Medicine*. *Time* listed Dr. Hákonarson's autism gene discovery project, reported in *Nature* in 2009, among the top 10 medical breakthroughs of that year. With more than 20 years of experience in pioneering genomics research and genome-wide mapping and association studies, Dr. Hákonarson has intimate knowledge of the complexities of large-scale genomics and drug development projects, and he has put together the necessary infrastructure and workflow processes to unravel these complexities for optimized deliverables of precision medicine programs.

**Josep Maria Haro Abad**

Research and Innovation Director
Saint John of God Health Park
Associate Professor
University of Barcelona
Spain

**Josep Maria Haro Abad, Ph.D.,** is the research and innovation director of Saint John of God Health Park in Barcelona, Spain, and associate professor of medicine at the University of Barcelona.

After his medical studies, he was trained in epidemiology and public health at the Johns Hopkins School of Hygiene and Public Health. Later he got his specialization in psychiatry at the Clinic Hospital of Barcelona. During the past 25 years he has worked both in clinical medicine and in public health research and has published more than 500 scientific papers. He has been included in the list of Clarivate Highly Cited Researchers in 2017 and 2018.

His areas of investigation have been epidemiology of mental disorders and the analysis of treatment outcomes. As a researcher in epidemiology of mental disorders, he has conducted studies on the prevalence and impact of disorders in the general population and the treatment of mental disorders in primary care. His last studies focus on the determinants of healthy aging, analyzing both the impact of physical and mental co-morbidity and the effect of societal and environmental aspects. In health outcomes, he has been interested in the consequences of mental disorders in patient functioning and quality of life, and the impact on society overall. During the last years, he has started research on the use of new ICT to evaluate and treat individuals with mental disorders.

Dr. Haro is principal investigator of one of the groups of the CIBERSAM network. In 2011 he received the award of best researcher from the Spanish Society of Biological Psychiatry. In 2018 he received the award of professional excellence of the Barcelona Medical Association. He is currently the European coordinator of the EU-funded project ATHLOS and SYNCHROS and was coordinator of the Roadmap for mental health and well-being research in Europe (ROAMER).

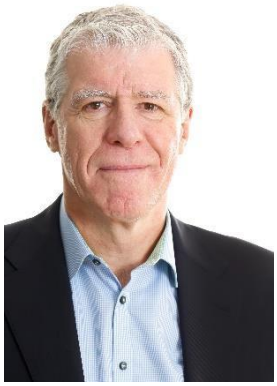
**Aroon Hingorani**

Director
UCL Institute of Cardiovascular Science
UK

**Professor Aroon Hingorani, M.A., Ph.D., FRCP, FESC,** is UCL Professor of Genetic Epidemiology. He is director of the UCL Institute of Cardiovascular Science, Cardiovascular Programme lead for the UCL Hospitals NIHR Biomedical Research Centre, and a co-investigator in the Precision Medicine Research Initiative of the HDRUK London site. He is a consultant physician at University College London Hospitals NHS Foundation Trust and an NIHR senior investigator.

He graduated in physiological sciences from the University of Oxford in in 1986 and in medicine, from the United Medical and Dental Schools (Guy's Hospital) in 1989. After clinical posts in London, he undertook research on the genetics of high blood pressure as an MRC Clinical Training Fellow in the University of Cambridge, obtaining his Ph.D. in 1997. He continued specialist training in general medicine and clinical pharmacology at UCL and UCL Hospitals, pursuing research on endothelial function in cardiovascular disease and then cardiovascular genomic, as a British Heart Foundation intermediate and then as senior fellow.

His current work focuses on the use of genetic studies in populations as a tool to identify and validate drug targets, using the Mendelian randomisation principle.



**Birgir Jakobsson**

Medical Advisor to the Minister of Health
Iceland

**Birgir Jakobsson M.D., Ph.D.**, graduated from the University of Iceland in 1975. He is a pediatrician by training; he earned his Ph.D. in pediatrics at the Karolinska institute in Stockholm in 1988 and became an associate professor in pediatrics at the Karolinska Institute in 1998. He became head of the department of pediatrics at Huddinge University Hospital in 1995 and chief of division at the same hospital in 1999. In 2003, he became a CEO of Capio St. Göran's Hospital in Stockholm and CEO for the Karolinska University Hospital in 2007, where he served until 2014. In 2015 he became chief medical officer (Surgeon General) of Iceland until retirement in April 2018. Dr. Jakobsson then became a medical advisor to the minister of health in Iceland.

**Elizabeth T. Jensen**

Assistant Professor
Wake Forest School of Medicine
USA

**Elizabeth T. Jensen, M.P.H, Ph.D**. is a reproductive, perinatal, pediatric epidemiologist with additional training in biomarker-based epidemiology and environmental exposures. Her research primarily focuses on etiologic factors in the development of pediatric chronic disease, including understanding factors contributing to disparities in health outcomes. Dr. Jensen earned both her M.P.H and Ph.D. from the University of North Carolina at Chapel Hill and completed a postdoctoral fellowship with the National Institute of Environmental Health Sciences (NIEHS). In addition to her appointment in Epidemiology and Prevention at Wake Forest School of Medicine, she has a joint appointment in the Department of Gastroenterology and holds an adjunct appointment at the University of North Carolina at Chapel Hill in the Department of Medicine, Division of Gastroenterology and Hepatology. Dr. Jensen has worked on numerous longitudinal cohort studies in pediatric populations, including analyses from the Mother-Child Cohort in Norway, the Collaborative Perinatal Project, the Extremely Low Gestational Age Newborns (ELGAN) Study, and primary data collection in the longitudinal SEARCH for Diabetes in Youth Study. Much of her research focuses on early life determinants for pediatric disease, including assessment of early life and environmental exposures in relation to pediatric health outcomes. Dr. Jensen leads an on-going study leveraging administrative databases and resources in Denmark to study early life environmental exposures in interaction with genotype in the development of eosinophilic esophagitis. She also leads a study of perfluorinated alkyl substances exposure in relation to metabolomic patterns in type 1 and type 2 diabetes through the Children's Health Exposure Assessment Resource (CHEAR), supported by NIEHS.

**Prabhat Jha**

Executive Director
Centre for Global Health Research
Canada

**Prabhat Jha, M.D., D.Phil.,** is an Endowed Professor in Global Health and Epidemiology at the University of Toronto and Canada Research Chair at the Dalla Lana School of Public Health and the founding director of the Centre for Global Health Research.

Prof. Jha is a lead investigator of the Million Death Study in India, which quantifies the causes of premature mortality in more than 2 million homes. His publications on tobacco control have enabled a global treaty now signed by more than 180 countries. He founded the Statistical Alliance for Vital Events, which focuses on reliable measurement of premature mortality worldwide.

Earlier, Prof. Jha served in senior roles at the World Health Organization and the World Bank. He was made an officer of the Order of Canada in 2012. Prof. Jha holds an M.D. from the University of Manitoba and a D.Phil. from Oxford University, where he studied as a Rhodes Scholar.

**Thomas Keane**

Team Leader
EGA and Archive Infrastructure
European Bioinformatics Institute (EBI)
UK

**Thomas Keane, M.Sc., Ph.D.,** joined EMBL-EBI as team leader in 2016. Prior to that, he led the Sequence Variation Infrastructure group in the Computational Genomics programme at the Wellcome Trust Sanger Institute. His interests are in using genomic technologies to learn about biological processes, with a particular focus on mouse and human disease.

**Barbara Kerstiëns**

Head of Unit
Health Directorate of the Directorate-General for Research and Innovation at the European Commission
Belgium

**Barbara Kerstiëns, M.D., M.P.H.**, is the head of the unit responsible for non-communicable diseases and the challenge of healthy aging in the Health Directorate of the Directorate-General for Research and Innovation at the European Commission. She has extensive experience in international public health, working for Médecins Sans Frontières, Johns Hopkins Bloomberg School of Public Health, and the Directorate-General for Development and Cooperation of the European Commission prior to joining DG Research and Innovation in 2012, where she has consistently worked in medical research and funding. Dr. Kerstiëns received her M.D. from the Katholieke Universiteit Leuven, a postgraduate certificate in tropical medicine from the Institute of Tropical Medicine in Antwerp, and her M.P.H. from Johns Hopkins Bloomberg School of Public Health.

**Gun Peggy Knudsen**

Executive Director, division of Health Data and Digitalisation
Norwegian Institute of Public Health
Norway

**Gun Peggy Knudsen, Ph.D.,** is the Executive Director of the division of Health Data and Digitalisation at the Norwegian Institute of Public Health (NIPH). The division is responsible for the management of five Norwegian mandatory population based health registries, several population based health studies and biobanks, where the Norwegian Mother and Child Cohort study with its 270.000 participants is the largest cohort study and biobank. Knudsen has a PhD in epigenetics, broad experience in managing research projects, both epidemiological and genetics projects, and have in depth knowledge of the cohort data, the biological materials and previous and ongoing genetics projects based on material at NIPH. The main research interest have been within the field of psychiatric genetics, and one of the current goals at the division is to complete the genotyping of all the participants in the Mother and Child Cohort study.

**Rongling Li**

Program Director and Epidemiologist
Division of Genomic Medicine
National Human Genome Research Institute (NHGRI)
National Institutes of Health (NIH)
USA

**Rongling Li, M.D., Ph.D., M.P.H.**, is a genetic epidemiologist and the program director for the International Hundred Thousand Cohort Consortium (IHCC). Prior to working on IHCC, she was the lead program director for the electronic Medical Records and Genomics (eMERGE) Network from 2009 to 2019. She has a long-standing interest in public health, especially genetic and genomic research on complex diseases and health-related phenotypes. As the lead program director for eMERGE, she was instrumental in facilitating the network's genomic discovery research efforts through both electronic phenotyping using electronic medical records (EMRs) and genotyping/sequencing using biorepository linked to those EMRs. Dr. Li has also been overseeing the application of genetic and genomic knowledge to clinical practice in the past 10 years.

Prior to joining the National Human Genome Research Institute in February 2009, Dr. Li worked at Rho, a biostatistics consulting firm/contract research organization (Chapel Hill, North Carolina), as an epidemiologist. She was also on the faculty of Morehouse School of Medicine and was a tenured professor at the University of Tennessee Health Science Center within the School of Medicine, Department of Preventive Medicine, and Center for Genomics and Bioinformatics.

Dr. Li has published more than 150 peer-reviewed journal articles. She was an associate editor of the *American Journal of Epidemiology* from 2006 to 2018.

**Daniel MacArthur**

Co-Director, Program in Medical and Population Genetics
Harvard University/Broad Institute/Massachusetts General Hospital
USA

**Daniel MacArthur, Ph.D.,** is an institute member at the Broad Institute of MIT and Harvard, and co-director of the Broad's Program in Medical and Population Genetics. In addition to his roles at the Broad, MacArthur is a group leader in the Analytic and Translational Genetics Unit at Massachusetts General Hospital and an assistant professor at Harvard Medical School. His work revolves around the use of large-scale genomic data to interpret genetic variants, particularly in the context of rare, severe genetic diseases.

MacArthur's team has assembled the largest collection of sequences of the protein-coding region (exome) of the human genome, creating a resource called the Genome Aggregation Database (gnomAD). This collection currently contains DNA sequencing data from over 140,000 individuals, and is made publicly available for anyone to use. As a result it has become the default reference database for clinical genetics labs, and is accessed over 15,000 times every day. It also serves as the basis for the Human Knockout Project, an ambitious global endeavor seeking to characterize the clinical impact of the disruption of each of the 20,000 genes in the human genome.

In addition, MacArthur leads a number of efforts applying genomic technologies to the diagnosis of very rare genetic diseases. He co-leads the Broad's Center for Mendelian Genomics, which uses both DNA and RNA sequencing technologies to investigate the genetic basis of rare diseases in thousands of families every year. In the first three years of the Center's operation it has been able to return genetic diagnoses to over 1,200 families – many of whom had been waiting many years for an answer using standard clinical testing – and identified nearly 100 likely new genes associated with a wide variety of diseases.

MacArthur was recognized for his work with the Harvard Medical School's Young Mentor Award in 2016, the Massachusetts General Hospital Martin Prize in 2017, and was also the first ever recipient of the American Society of Human Genetics Early-Career Award in 2017.

MacArthur completed his Ph.D. at the Institute for Neuromuscular Research in Sydney, Australia, where he studied a loss-of-function variant in the human *ACTN3* gene associated with variation in muscle strength and athletic performance. He later served as a postdoctoral fellow at the Wellcome Trust Sanger Institute in Hinxton, UK, where he led the annotation of gene-disrupting ("loss-of-function") variants as part of the 1000 Genomes Project Consortium.

Michaela Th. Mayrhofer

Chief Policy Officer CS ELSI/Chief Coordinator Officer
The Biobanking and Biomolecular resources Research Infrastructure (BBMRI)-European Research Infrastructure Consortium (ERIC)
Austria

**Michaela Th. Mayrhofer, Ph.D.,** earned her Ph.D. from both the École des Hautes Études en Sciences Sociales and the University of Vienna; her work was shortlisted by the Austrian Society for Political Science for "best thesis 2010." During her academic career, she has been a fellow at renowned research institutions in Belgium, the United Kingdom, France, and Switzerland. Today she serves as BBMRI-ERIC's chief coordination and policy officer, coordinates the BBMRI-ERIC Common Service ELSI, leads the ELSI Work Packages for several EU projects, and spearheads the GDPR Code of Conduct for Health Research Initiative.

**Nicola Mulder**

Head
Computational Biology Division at the University of Cape Town
South Africa

**Nicola Mulder, Ph.D.,** heads the Computational Biology Division at the University of Cape Town and leads H3ABioNet, a pan African bioinformatics network of 28 institutions in 16 African countries. H3ABioNet is developing bioinformatics capacity to enable genomic data analysis on the continent. Prior to her position at the University of Cape Town (UCT), she worked for more than 8 years at the European Bioinformatics Institute in Cambridge, as a team leader for bioinformatics resources. At UCT her research focuses on genetic determinants of susceptibility to disease, African genome variation, microbiomes, genomics, and infectious diseases from the host and pathogen perspectives. Her group provides bioinformatics services for local researchers, through which they develop visualization and analysis tools for high-throughput biology. Her team has also developed new and improved algorithms for the analysis of African genetic data and for downstream analysis and interpretation of GWAS data. Prof. Mulder is actively involved in training and education, including bioinformatics curriculum development. She co-chairs international committees on bioinformatics education and sits on a number of scientific advisory boards for African and international projects.

**Kenny Nguyen**

Biological Informatics Scientist
Center for Applied Genomics
Children's Hospital of Philadelphia
USA

**Kenny Nguyen, Ph.D.,** works as a biological informatics scientist at Children's Hospital of Philadelphia's Center for Applied Genomics. He was trained in computational and theoretical chemistry using simulation methods from all-atom and coarse-grain molecular dynamics and force field development—based on the principles of statistical thermodynamics—to render self-assembly. He also has complementary background in computational quantum chemistry, with expertise in torsional potential energies from steric (geometric) and electronic interactions of conjugated dienes. He has transitioned to a more traditional approach in computational structural biology by investigating the conformation of macromolecules, as well as those bound to therapeutic targets. He now uses this knowledge—using high-performance and throughput computing coupled with advanced/enhanced modeling methods—to elucidate information (structure, dynamics, energetics, and function) regarding gene–drug (and protein–ligand) interactions related to the overlap of pharmacology and genetics/genomics. These techniques are also used to determine therapeutics through computer-aided drug design for repurposing, repositioning, and rescue. More recently, a quantitative systems approach and deep machine learning methods in artificial intelligence are being explored for analyses in rare/orphan and other complex diseases.

**Anthony Philippakis**

Cardiologist
Brigham and Women's Hospital
Venture Partner, GV
Chief Data Officer
Broad Institute of Harvard and MIT
USA

**Anthony Philippakis, M.D., Ph.D.,** is a physician, geneticist, and data scientist. He is currently a cardiologist at Brigham and Women's Hospital, a venture partner at GV, and the chief data officer at Broad Institute of Harvard and MIT.

Dr. Philippakis studied mathematics as an undergraduate at Yale University, followed by getting a master's in mathematics at Cambridge University. He completed an M.D. at Harvard Medical School and a Ph.D. in biophysics at Harvard, working to develop computational methods for understanding transcriptional regulation. He completed his medical residency and cardiology fellowship at Brigham and Women's Hospital.

Dr. Philippakis is committed to bringing genome sequencing and data science into the practice of clinical medicine. As a clinician, he specializes in the care of patients with rare genetic cardiovascular diseases. He co-chairs the Scientific Advisory Board of Global Genes, is a strategic adviser to the American Heart Association, and sits on the Clinical Working Group of the Global Alliance for Genomics and Health.

**Gad Rennert**

Director
Clalit National Israeli Cancer Control Center (NICCC) and National Personalized Medicine Program
Israel

**Gad Rennert, M.D., Ph.D.,** has been chairman of the Carmel Medical Center Department of Community Medicine and Epidemiology since 1992. He is a professor and the head of the public health and epidemiology teaching group at the Technion Faculty of Medicine.

Professor Rennert is also Director of the National Israeli Cancer Control Center and the Department of Epidemiology and Disease Prevention of Clalit and is leading its National Personalized Medicine Program offering testing, advice and policy on individualized molecular testing which dictates cancer risk and suitability for cancer treatments. He is responsible for the national breast and colorectal cancer detection programs in Israel and is a member of the National Oncology Council.

In 1984, Professor Rennert received his medical degree from Ben-Gurion Medical School. He received his PhD in Public Health from the University of North Carolina. He focuses his studies on understanding the behavioral and biological causes of cancer, with special emphasis on gene-environment interactions. He has been an invited speaker in key conferences, such as the Personalized Medicine World Conference, UPCP, American Society of Clinical Oncology, American Association of Cancer Research, St. Galen Cancer Prevention conference and San Antonio Breast Cancer Symposium.

In addition to his activities at the Technion, Dr. Rennert is a reviewer for more than 30 international journals, an associate editor of two and serves on 10 editorial boards. He has published more than 200 papers in leading journals such as the NEJM, Science and Nature.

**Laura Lyman Rodriguez**

Director
Division of Policy, Communications, and Education
National Human Genome Research Institute (NHGRI)
National Institutes of Health (NIH)
USA

**Laura Rodriguez, Ph.D.,** is the director of the Division of Policy, Communications, and Education. In this capacity, she works to develop and implement policy for research initiatives at NHGRI and NIH, design communication and outreach strategies to engage the public in genomic science, and prepare health care professionals for the integration of genomic medicine into clinical care. Dr. Rodriguez is particularly interested in the policy and ethics questions related to the inclusion of human research participants in genomics and genetics research.

Dr. Rodriguez's career has included positions in the legislative, advocacy, and non-governmental policy arenas, where she focused on a range of topics, including research ethics, intellectual property, and human research participant regulations.

**Patrick Sleiman**

Associate Director
Center for Applied Genomics (CAG) at the Children's Hospital of Philadelphia (CHOP)
USA

**Patrick Sleiman, Ph.D.** is associate director of the Center for Applied Genomics (CAG) at the Children's Hospital of Philadelphia (CHOP) and assistant professor in the department of pediatrics of the University of Pennsylvanian Perelman School of Medicine (PSOM). His primary research interests are in uncovering the genetic basis of human traits and diseases, extending the application of genetic data to healthcare and identifying novel genetically-informed therapeutics. He earned his PhD in genetics from the Galton Laboratories at University College London and completed postdoctoral training at the Institute of Neurology, University of London.

**Eleftheria Zeggini**

Director
Institute of Translational Genomics
Helmholtz Zentrum München
Germany

**Eleftheria Zeggini, Ph.D.,** obtained a B.Sc. in biochemistry from the University of Manchester Institute of Science and Technology (UMIST) in 1999 and a Ph.D. in immunogenetics of juvenile arthritis from the arc Epidemiology Unit, University of Manchester, in 2003. She then undertook a brief statistical genetics postdoc focusing on rheumatic disorders, at the Centre for Integrated Genomic and Medical Research, University of Manchester, before moving to the Wellcome Trust Centre for Human Genetics, University of Oxford, to work on the genetics of type 2 diabetes. In 2006, Dr. Zeggini was awarded a Wellcome Trust Research Career Development Fellowship to examine design, analysis, and interpretation issues in large-scale association studies. She joined the Wellcome Sanger Institute faculty in November 2008, where she led the Analytical Genomics of Complex Traits group for 10 years. In September 2018, she moved to Helmholtz Munich to establish a new Institute of Translational Genomics.

*Note that the following bios and/or photos were retrieved from online sources: Daniel MacArthur and Gad Rennert.*

# Meeting Venue & Hotel


**Grand Hotel Reykjavík**

Sigtún 38
105 Reykjavík, Iceland
+354 514 8000

Grand Hotel Reykjavík is a four-star superior hotel for business travelers, conference guests and tourists who demand excellent service and facilities. Located a quick 15-minute ride from the center of downtown Reykjavík and 40 minutes from Keflavík International Airport, the hotel has one restaurant, The Grand Brasserie, that offers a variety of Icelandic and Nordic dishes with a wide range of wines and cocktails. There is also a spa in the hotel.

Grand Hotel Reykjavík conforms to the Nordic Ecolabelling criteria for hotels. This guarantees that the highest standards regarding environmental measures, health, functionality, and quality requirements have been met.

Hotel Website

### Getting to the Hotel

Grand Hotel Reykjavík is approximately a 40-minute car ride from the Keflavík International Airport. We suggest taking a taxi to get to the hotel; taxis should be available at the airport upon arrival.

# Sponsor

A big thank-you to the sponsor of the 2nd International Cohorts Summit for their generous contribution and continued support!

# Participant List


**Malak Abedalthagafi, M.D.**
King Abdulaziz City for Science and Technology
Saudi Arabia
malthagafi@kacst.edu.sa

**Jesus Alegre-Díaz, M.D.**
National Autonomous University of Mexico
Mexico
inypcjad@gmail.com

**Nahla Afifi, Ph.D.**
Qatar Biobank-Qatar Foundation
Qatar
nafifi@qf.org.qa

**Jessica Alföldi, Ph.D.**
Broad Institute
USA
jalfoldi@broadinstitute.org

**Garnet Anderson, Ph.D.**
Fred Hutchinson Cancer Research Center
USA
garnet@whi.org

**Philip Awadalla, Ph.D.**
Canadian Partnership for Tomorrow Project
Canada
philip.awadalla@oicr.on.ca

**Sarah Bauermeister, Ph.D.**
European Prevention of Alzheimer's Dementia (EPAD)/Cognitive Neuropsychologist
Dementias Platform UK (DPUK)
UK
sarah.bauermeister@psych.ox.ac.uk

**Ewan Birney, Ph.D.**
The European Bioinformatics Institute
UK
birney@ebi.ac.uk

**Dan Brake, M.I.T.E.**
Sequence Bioinformatics
Canada
dan@sequencebio.com

**Adam Butterworth, Ph.D.**
University of Cambridge
UK
asb38@medschl.cam.ac.uk

**Lon Cardon, Ph.D.**
BioMarin Pharmaceutical Inc.
USA
lon.cardon@bmrn.com

**Juan P. Casas, M.D., Ph.D.**
VA Boston Healthcare System
USA
juan.casasromero@va.gov

**John Chambers, Ph.D.**
Nanyang Technological University
Singapore
john.chambers@ic.ac.uk

**Zhengming Chen, D.Phil.**
Oxford University
UK
zhengming.chen@ndph.ox.ac.uk

**Tammy Clifford, Ph.D., M.Sc.**
Canadian Institutes of Health Research
Canada
tammy.clifford@cihr-irsc.gc.ca

**Rory Collins**
Oxford University
UK
rory.collins@ndph.ox.ac.uk

**John Connolly, Ph.D.**
Children's Hospital of Philadelphia
USA
connollyj1@chop.edu

**Nancy Cox, Ph.D.**
Vanderbilt University
USA
nancy.j.cox@vanderbilt.edu

**Mark Daly, Ph.D.**
Institute for Molecular Medicine Finland
Finland
mark.daly@helsinki.fi

**Mary De Silva, Ph.D., M.Sc.**
Wellcome Trust
UK
m.desilva@wellcome.ac.uk

**Joshua Denny, M.D., M.S.**
Vanderbilt University
USA
josh.denny@vanderbilt.edu

**Emanuele Di Angelantonio, Ph.D., M.Sc.**
University of Cambridge
UK
ed303@medschl.cam.ac.uk

**Rajesh Dikshit, Ph.D.**
Tata Memorial Hospital
India
dixr24@hotmail.com

**Sylvain Durrleman, M.D.**
INSERM/Aviesan Institute of Public Health
France
sylvain.durrleman@gmail.com

**Mark Effingham, Ph.D.**
UK Biobank
UK
mark.effingham@ukbiobank.ac.uk

**Margaret Ehm, Ph.D.**
GSK
USA
meg.g.ehm@gsk.com

**Robert Eiss, M.A.**
John E. Fogarty International Center
USA
robert.eiss@nih.gov

**Heather Eliassen, Sc.D., S.M.**
Harvard T.H. Chan School of Public Health,
Harvard University
USA
heliasse@hsph.harvard.edu

**Paul Elliott, Ph.D.**
Imperial College London
UK
p.elliott@imperial.ac.uk

**Jonathan Emberson, Ph.D., M.Sc.**
University of Oxford
UK
jonathan.emberson@ndph.ox.ac.uk

**Arash Etemadi, M.D., Ph.D.**
National Cancer Institute
USA
arash.etemadi@nih.gov

**Catterina Ferreccio, M.D., M.S.P.**
Pontificia Universidad Católica de Chile
Chile
catferre@gmail.com

**Tom Fowler**
Genomics England
UK
claire.locatelli-bolton@genomicsengland.co.uk

**Neal Freedman, Ph.D., M.P.H.**
National Cancer Institute
USA
freedmanne@nci.nih.gov

**Nobuo Fuse**
Tohoku University Tohoku Medical Megabank
Organization
Japan
fusen@megabank.tohoku.ac.jp

**John Gallacher, Ph.D.**
University of Oxford/Dementias Platform UK
UK
john.gallacher@psych.ox.ac.uk

**J. Michael Gaziano**
VA Boston Healthcare System
USA
jmgaziano@bwh.harvard.edu

**Kelly Gebo, M.D., M.P.H.**
National Institutes of Health, *AoU*
USA
kelly.gebo@nih.gov

**Christian Gieger, Ph.D.**
Helmholtz Zentrum München
Germany
christian.gieger@helmholtz-muenchen.de

**Geoffrey Ginsburg, M.D., Ph.D.**
Duke University
USA
geoffrey.ginsburg@duke.edu

**Roger Glass, M.D., Ph.D.**
John E. Fogarty International Center
USA
glassr@mail.nih.gov

**Marcel Goldberg, M.D., Ph.D.**
INSERM UMS 11
France
marcel.goldberg@inserm.fr

**Peter Goodhand**
Global Alliance for Genomics & Health
Canada
peter.goodhand@ga4gh.org

**Eric Green, M.D., Ph.D.**
National Human Genome Research Institute
USA
egreen@nhgri.nih.gov

**Fran Grodstein, Sc.D.**
Brigham and Women's Hospital
USA
fran.grodstein@channing.harvard.edu

**Joseph Grzymski, Ph.D.**
Renown Health/Desert Research Institute
USA
joeg@dri.edu

**Hákon Hákonarson, M.D., Ph.D.**
Children's Hospital of Philadelphia
USA
hakonarson@email.chop.edu

**Josep Maria Haro, Ph.D.**
Parc Sanitari Sant Joan de Déu
Spain
jmharo@pssjd.org

**Aroon Hingorani, Ph.D.**
University College London
UK
a.hingorani@ucl.ac.uk

**Atsushi Hozawa, M.D., Ph.D.**
Tohoku University Tohoku Medical Megabank Organization
Japan
hozawa@megabank.tohoku.ac.jp

**Birgir Jakobsson, M.D., Ph.D.**
Ministry of Health
Iceland
birgir.jakobsson@hrn.is

**Rahman Jamal, Ph.D.**
Universiti Kebangsaan Malaysia
Malaysia
rahmanj@ppukm.ukm.edu.my

**Sun Ha Jee, Ph.D., M.P.H.**
Yonsei University
South Korea
jsunha1066@gmail.com

**Yon Ho Jee, M.Sc., S.M.**
Yonsei University
South Korea
minniejee93@gmail.com

**Elizabeth Jensen, Ph.D., M.P.H.**
Wake Forest School of Medicine
USA
ejensen@wakehealth.edu

**Jae-Pil Jeon, Ph.D.**
Korea National Institute of Health
South Korea
jaepiljeon@hanmail.net

**Prabhat Jha, M.D., D.Phil.**
University of Toronto
Canada
prabhat.jha@utoronto.ca

**Guðni Thorlacius Jóhannesson, Ph.D.**
President of Iceland
Iceland

**Farin Kamangar, M.D., Ph.D.**
Morgan State University
USA
farin.kamangar@gmail.com

**Norihiro Kato, M.D., D.Phil.**
National Center for Global Health and Medicine
Japan
nokato@hosp.ncgm.go.jp

**Thomas Keane, M.Sc., Ph.D.**
European Bioinformatics Institute
UK
tk2@ebi.ac.uk

**Barbara Kerstiëns, M.D., M.P.H.**
European Commission, Directorate-General for Research and Innovation
Belgium
barbara.kerstiens@ec.europa.eu

**Sung Soo Kim, Ph.D., M.P.H.**
Korea National Institute of Health
South Korea
ksungsoo@korea.kr

**Gun Peggy Knudsen, Ph.D.**
Norwegian Insititute of Public Health
Norway
gun.peggy.knudsen@fhi.no

**David Ledbetter, Ph.D.**
Geisinger Health
USA
dhledbetter@geisinger.edu

**Sarah Lewington, D.Phil., M.Sc.**
University of Oxford
UK
sarah.lewington@ndph.ox.ac.uk

**Rongling Li, M.D., Ph.D., M.P.H.**
National Human Genome Research Institute
USA
lir2@mail.nih.gov

**Chi-Ming Liang, Ph.D.**
Academia Sinica
Taiwan
cmliang@gate.sinica.edu.tw

**Rachel Liao, Ph.D.**
Broad Institute
USA
rliao@broadinstitute.org

**Paulo Lotufo, M.D., Dr.P.H.**
University of São Paulo
Brazil
palotufo@usp.br

**Beatrice Lucaroni**
European Commission, Directorate-General for Research and Innovation
Belgium
beatrice.lucaroni@ec.europa.eu

**Chris Lunt**
National Institutes of Health
USA
chris.lunt@nih.gov

**Daniel MacArthur, Ph.D.**
Harvard University/Broad Institute/Massachusetts General Hospital
USA
danmac@broadinstitute.org

**Per Magnus, M.D., Ph.D.**
Norwegian Institute of Public Health
Norway
per.magnus@fhi.no

**Reza Malekzadeh, M.D.**
Tehran University of Medical Sciences
Iran
dr.reza.malekzadeh@gmail.com

**Teri Manolio, M.D., Ph.D.**
National Human Genome Research Institute
USA
manoliot@mail.nih.gov

**Tohru Masui**
Keio University
Japan
tmasui@keio.jp

**Prashant Mathur**
Indian Council of Medical Research
India
mathurp.hq@icmr.gov.in

**Michaela Th. Mayrhofer, Ph.D.**
BBMRI-ERIC
Austria
michaela.th.mayrhofer@bbmri-eric.eu

**Hamdi Mbarek, Ph.D.**
Qatar Genome
Qatar
hmbarek@qf.org.qa

**Joe McNamara, Ph.D.**
Medical Research Council
UK
joe.mcnamara@mrc.ukri.org

**Martin McNamara, Ph.D., M.P.H.**
Sax Institute
Australia
martin.mcnamara@saxinstitute.org.au

**Mads Melbye**
Statens Serum Institut
Denmark
mme@ssi.dk

**Usha Menon, Ph.D., RN**
University College London
UK
u.menon@ucl.ac.uk

**Andres Metspalu, M.D., Ph.D.**
University of Tartu
Estonia
andres.metspalu@ut.ee

**Takayuki Morisaki, Ph.D.**
University of Tokyo
Japan
morisaki@ims.u-tokyo.ac.jp

**Nicola Mulder, Ph.D.**
University of Cape Town
South Africa
nicola.mulder@uct.ac.za

**Yoshinori Murakami, M.D., Ph.D.**
BioBank Japan/ University of Tokyo, Institute of Medical Science
Japan
ymurakam@ims.u-tokyo.ac.jp

**Matthew Nelson, Ph.D.**
GlaxoSmithKline
UK
matthew.r.nelson@gsk.com

**Kenny Nguyen**
Children's Hospital of Philadelphia
USA
nguyenk6@chop.edu

**Thea Norman, Ph.D.**
Bill & Melinda Gates Foundation
USA
thea.norman@gatesfoundation.org

**Donna Parker, M.P.H.**
Duke University
USA
donna.l.parker@duke.edu

**Alexandre Pereira, M.D., Ph.D.**
University of São Paulo
Brazil
alexandre.pereira@incor.usp.br

**Mauro Petrillo, M.Sc.**
European Commission
Belgium
mauro.petrillo@ec.europa.eu

**Anthony Philippakis, M.D., Ph.D.**
Broad Institute
USA
aphilipp@broadinstitute.org

**Brittany Ploss, M.Sc.-G.H.**
Duke University/Global Genomic Medicine Collaborative
USA
brittany.zick@duke.edu

**Erica Pufall, Ph.D., M.Sc.**
Wellcome Trust
UK
e.pufall@wellcome.ac.uk

**Teji Rakhra-Burris**
Global Genomic Medicine Collaborative
USA
teji@g2mc.org

**Gad Rennert, M.D., Ph.D.**
Carmel Medical Center/Technion–Israel Institute of Technology
Israel
rennert@technion.ac.il

**Gabriela Repetto**
Clínica Alemana Universidad del Desarrollo
Chile
grepetto@udd.cl

**Jessica Reusch, Ph.D.**
National Human Genome Research Institute
USA
jessica.reusch@nih.gov

**Laura Lyman Rodriguez, Ph.D.**
National Human Genome Research Institute
USA
laura.rodriguez@nih.gov

**Norie Sawada**
National Cancer Center Japan
Japan
nsawada@ncc.go.jp

**Catherine Schaefer, Ph.D.**
Kaiser Permanente
USA
cathy.schaefer@kp.org

**Alan Shuldiner, M.D.**
Regeneron Genetics Center
USA
alan.shuldiner@regeneron.com

**Patrick Sleiman, Ph.D.**
Children's Hospital of Philadelphia
USA
patricksleiman@me.com

**Yuriko Suzuki, M.D., Ph.D., M.P.H.**
Japan Agency for Medical Research and Development (AMED)
Japan
yuriko-suzuki@amed.go.jp

**Anthony Swerdlow**
Institute of Cancer Research
UK
anthony.swerdlow@icr.ac.uk

**Patrick Tan, M.D., Ph.D.**
Agency for Science, Technology and Research
Singapore
patrick_tan@a-star.edu.sg

**David van Heel, D.Phil., M.A.**
Queen Mary University of London
UK
d.vanheel@qmul.ac.uk

**Mara Vitolins, Dr.P.H., M.P.H.**
Wake Forest University
USA
mvitolin@wakehealth.edu

**Andreas Weser**
Norwegian University of Science and Technology
USA
andreas.weser@ntnu.no

**Tsungfu Yu, Ph.D.**
Academia Sinica
Taiwan
albertyu@ibms.sinica.edu.tw

**Eleftheria Zeggini**
Institute of Translational Genomics
Germany
eleftheria.zeggini@helmholtz-muenchen.de

**Wei Zheng, M.D., Ph.D.**
Vanderbilt University
USA
wei.zheng@vanderbilt.edu

**Marie Zins**
INSERM UMS 11
France
marie.zins@inserm.fr

# Participating Cohorts

- 45 and Up Study
- Africa Centre for Health and Population Studies
- *All of Us* Research Program
- Airwave Health Monitoring Study
- Barshi Cohort
- BBMRI-ERIC Colon Cancer Cohort
- BioBank Japan
- BioVU
- Brazilian Longitudinal Study of Adult Health (ELSA-Brasil)
- Chinese Newborn Sequencing Project
- Children's Hospital of Philadelphia
- China Kadoorie Biobank
- CONSTANCES
- Danish National Biobank
- Dementias Platform UK
- East London Genes and Health
- Environmental Influences on Child Health Outcomes (ECHO)
- Estonian Biobank
- Finnish Genome Project (FinnGen)
- Generations Study
- German National Cohort (NAKO)
- Golestan Cohort Study
- Prospective Epidemiological Research Studies in IrAN (PERSIAN) Cohort
- Healthy Nevada
- Human Heredity and Health in Africa (H3Africa)
- Israel Genome Project
- Japan Multi-Institutional Collaboration Cohort Study
- Japan Public Health Center-based Prospective Study
- Japan Public Health Center-based Prospective Study for the Next Generation
- Kaiser Permanente Research Program on Genes, Environment and Health
- Korean Cancer Prevention Study-II
- Korean Biobank Project
- Korean Genome and Epidemiology Study
- Malaysian Cohort
- Maule Cohort (MAUCO Study)
- Mexico City Prospective Study
- Million Veteran Program
- MyCode Community Health Initiative
- Newfoundland 100K Genome Project
- SG100K
- Nurses' Health Study
- Nurses' Health Study II
- Nord-Trondelag Health Study (HUNT)
- Norweigan Mother and Chilld Cohort Study (MoBa)
- Ontario Health Study
- Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO)
- Qatar Biobank
- Qatar Genome Project
- Saudi Human Genome Project
- Shanghai Men and Women's Health Studies
- Singapore National Precision Medicine Program (SG100K)
- South Asia Biobank
- Taiwan Biobank
- Tohoku Medical Megabank
- UK Biobank
- UK Blood Donor Cohorts
- UK Collaborative Trial of Ovarian Cancer Screening
- UK Longitudinal Women's Study
- Women's Health Initiative

Global Genomic
Medicine Collaborative

# IHCC Request for Information (RFI) Submissions

International Cohorts Summit

# A Global Understanding of the Role of Proteomic and Metabolomic Profiles in Health and Disease

**Authors:** Francine Grodstein, Sc.D.; Heather Eliassen, Sc.D.

## 1. Idea Summary

Circulating protein concentrations may vary as a function of age, health status, and environmental exposure. There are several well-established protein biomarkers of disease risk including prostate-specific antigen for prostate cancer, hemoglobin A1c for diabetes, and anti-citrullinated protein antibodies for rheumatoid arthritis.

Additionally, metabolomics, as the end-product of the other "omes", reflects both endogenous processes as well as exogenous exposures. Dysregulated metabolism has been shown to play a role in metabolic diseases, and epidemiologic evidence supports a role in other chronic diseases as well. Exploring broader proteomic and metabolomic profiles across diverse populations will allow for a more thorough understanding of the role of these biomarkers in health and disease. Leveraging the unique diversity of cohorts in the IHCC, we propose to investigate differences in proteomic and metabolomic profiles across populations that span a breadth of racial/ethnic, geographic, and socioeconomic representation. We propose to investigate how lifestyle factors influence these profiles, and to examine the role of these omics profiles in the development of chronic disease. Proteomic and metabolomic profiling on plasma or serum samples will be conducted across participating cohorts, pre-specified analyses would be carried out within cohorts, and results would be meta-analyzed across IHCC cohorts.

## 2. Structured Abstract

Background and rationale: In the progression from genotype to phenotype, proteomics and metabolomics integrate both genetic and environmental impacts on the body's systems and represent an opportunity to investigate the influence of this intersection on human health. Methods to measure proteomics, the set of proteins produced within an organism, and metabolomics, the measurement of small molecule intermediates and products of metabolism, have been established to capture hundreds to thousands of markers. There is compelling evidence from current epidemiologic studies to support a role of metabolism in the development of cardiometabolic diseases, but also emerging evidence to support a role in the etiology of several types of cancer. Further, proteomics profiling includes sets of proteins important in inflammation and immune function processes that play a critical role in several chronic diseases.

Idea: We propose to measure circulating proteomics and metabolomics in archived blood samples using targeted platforms to measure identified proteins and metabolites. By leveraging archived blood samples available across cohorts in the IHCC, across multiple countries and continents, we can investigate differences in proteomic and metabolomic profiles by country, race/ethnicity, sex, and age. We propose to examine the role of lifestyle factors, with some markedly different profiles across the globe, in these omic profiles. For example, understanding the roles smoking, diet, physical activity, obesity, education, other socio-economic factors, or underlying cardiometabolic diseases play in determining proteomic and metabolomic profiles will enhance our understanding of how these risk factors may contribute to chronic disease and mortality. Finally, as follow-up after blood collections continues in cohorts, we can begin to compare these omics profiles in incident cases versus controls selected on the basis of specific chronic diseases (i.e., cancers, cardiovascular diseases, cerebrovascular diseases).

Impact: Better understanding the role proteomics and metabolomics play in health and disease could reveal novel mechanisms underlying risk factors, and lead to identification of interventions that could impact health worldwide.

Timeline: The work involved in this proposed idea would include: 1) metabolomics and proteomics assays in archived plasma or serum samples within each participating cohort; 2) detailed analysis plans

to be carried out within each cohort, including cross-sectional analyses of lifestyle and environmental factors and proteomic/metabolomic profiles and case-control analyses of disease; 3) meta-analysis of results across cohorts. We anticipate the proteomic and metabolomic profiling would be completed in years 1 and 2, with concurrent development of analytic plans. Analysis of individual cohort data would be conducted in year 2 and the start of year 3, with meta-analysis across cohorts conducted in year 3.

Estimated funding required: Proteomics assay (with a commercial platform e.g., Olink or SOMAscan) and metabolomics assay (e.g., Metabolon or Biocrates) would be required at a cost of $1000-1500 per sample. Funding required for individual cohort needs would vary depending on infrastructure and cohort size, but would need to cover accessing and shipping blood samples, data management to conduct QC pipeline and incorporate -omics results with other cohort data, and analytic programming to conduct analyses. Additional funding may be required to cover the time of investigators coordinating the efforts within individual cohorts as well as investigator time to carry out this consortium project.

### 3. Additional Information

It is clear the downstream implications of the integration of genetics and environment play a role in health and disease development, and the incorporation of cohorts that capture ethnic/racial diversity as well as geographical and environmental diversity would allow for a large contribution to the knowledge of the roles proteomics and metabolomics play in health and disease.

Plasma or serum samples will be required from participating cohorts. Required data would include lifestyle, environmental (including geocoding) and disease factors of interest (e.g., smoking status and history, BMI (at points across the life course if available), disease history), as well as incidence of diseases of interest.

While some cohorts may have undertaken smaller-scale metabolomics projects, we assume most cohorts will need to supplement metabolomics profiling on most cohort samples, and proteomics is likely not as widespread and will need to be completed de novo for this project.

To achieve the near-term accomplishment, we anticipate utilizing a meta-analysis approach for this project, designing analyses to be carried out within individual cohorts, and meta-analyzing across cohorts. Longer-term approaches could include incorporation and harmonization of lifestyle and biomarker data across cohorts to conduct pooled analyses.

# A Precision Medicine Approach to Multi-morbidity in Cardio-metabolic Disease and Dementia

**Author:** John Gallacher, Ph.D.

## Background

This proposal is designed to identify and address the challenges of accessing data from multiple large-scale cohorts. The use case is to investigate the impact of multi-morbidity on cardio-metabolic disease and dementia. Due to size, breadth and depth of measurement IHCC cohorts provide the opportunity to investigate these associations at molecular, system, and clinical levels.

The proposal is to offer the DPUK Data Portal as a trusted third-party multi-cohort data management platform that would be free at the point of access. In the context of this project the proposal is to provide a relatively open research environment within which bona-fide scientists can gather and collaborate around a specific research question “multi-morbidity, cardio-metabolic disease and dementia”. The data portal provides a secure and fully auditable environment for cross-cohort analyses. All analyses must be conducted remotely as data, once uploaded to the Portal cannot be downloaded by third party researchers.

## Research question

How can analysis of large-scale “omic” data, survey data, and EHR-linked data inform:

- disease aetiology and sub-classification
- disease risk prediction
- therapeutic target prioritisation
- fundamental biological insight

## Methods

IHCC cohorts will be scoped to establish the relevant datasets and their degree of overlap.

Both traditional regression methods and machine learning techniques will be used to identify and interpret clusters and patterns of association between molecular, survey, and clinical data.

The user experience of developing and accessing these data will be used to develop data discovery and visualisation tools suitable for use with large-scale cohorts.

The work will be conducted as a collaborative involving the relevant cohorts and other invited investigators. The project provides the opportunity for a broad-based invitation to exploit the data corpus established by this collaboration.

## Strategic fit

This proposal supports the IHCC goal of increasing the realised scientific value of largescale datasets.

## Likely Impact

The proposal is intended to have impact in the near term (24 months) by:

- Providing a unified framework and approaches/tools for the combined study of omics, survey and linkage data at scale, thereby serving as the nucleus for an ongoing programme of data discovery and access development.
- Yielding new insights into the causes, subtypes, outlook and potential treatments for important conditions, resulting in publications; and
- Shaping the priorities and measurements in major cohorts which are contemplating omic assays, detailed phenotyping, and record linkage.

## Infrastructure

The DPUK Data Portal https://portal.dementiasplatform.uk/ has three core utilities: data discovery, access, and analysis. Of particular interest here is data analysis. Once approval has been granted by a cohort research team and a data access agreement has been completed, data are made available within the secure analysis area of the portal. This analysis area includes the use of several widely used general statistical packages (R, STATA, SPSS, SAS, Matlab, Python). Specialist software can be made available on request and bespoke software can be uploaded upon approval.

Researchers are provided with a personal virtual desktop infrastructure which requires two-factor authentication to access. The standard desktop specification is optimised for epidemiologic survey data and includes 8 GB RAM and 4 CPUs which is sufficient for most analyses. Bespoke desktop configurations can be requested for computationally intensive operations.

The Data Portal allows joint use of data within research consortia. Although the virtual desktop is the researcher's own personal virtual laboratory, for distributed research groups and for consortia, the portal can be used to hold a core dataset which can then be accessed by researchers in multiple locations without risk to the integrity of the core. This network of virtual desktops is flexible and can be configured in terms of access rights and capacity according to the requirements of the consortium. For the purposes of this proposal, collaborating IHCC cohorts would be considered to be the consortium and they would establish the rules for access

**Strengths**

This proposal has value at several levels:

- It will provide valuable evidence on a key scientific question that affects clinical decision making at many levels
- It will identify gaps in the data corpus of IHCC cohorts and inform how best to enhance IHCC cohorts
- It will help IHCC cohorts establish actual legal, professional and practical constraints on managing data access according to data-type and jurisdiction
- It will inform IHCC on the practical and cultural challenges involved in developing data repositories
- It will inform IHCC on the development of data discovery tools for IHCC cohorts

**Challenges and risks**

Challenges include:

- The Data Portal was not designed to manage the potential intensity of traffic that this project might generate. Agreement is required on the levels of performance that IHCC require and how this may be achieved
- Analysts are not used to accessing data remotely and will need to be assisted so that accessing data remotely becomes a routine procedure.
- Individual cohort access procedures will be bespoke. Flexibility is required to develop workable solutions. For example, where cohort data may not be uploaded beyond a national boundary, federated analysis approaches will be required.

A major risk is that IHCC cohorts may not wish to support the project which has not been initiated by a cohort research team. Although the DPUK is well placed to contribute scientifically to the project in terms of analytic and technical resource, DPUK is very happy not to lead this scientifically if that would be helpful, and would prefer not to be the governance lead. The key issue here is gaining the support of the cohort research teams. One solution might be to offer leadership of specific analyses to leaders in the field from collaborating cohorts.

Other risks are low:

- The technology solution is established. The Data Portal lies within the SAIL environment which operates to ISO 27001.
- Any technology development is a matter of capacity and speed, not build or governance
- There is no change of legal status required for any dataset

- There is no change in data governance principles required for any dataset
- There is no loss of scientific attribution for any collaborating research team

**Illustrative project plan** (Months)
Pre-project activity (on the basis of a pre-award letter):

1. Hire staff (-3-0)
2. Enhance the Data Portal (-3-0)
3. Establish the consortium governance and communications structures (-3-0)

Scientific activity

4. Conduct a gap analysis of the data available from IHCC cohorts (1-6)
5. Prioritise cohorts for enrichment according to informativeness for the hypothesis and wider scientific value (4-6)
6. Invite relevant IHCC cohort to join the collaboration (7)
7. Establish the data access procedures and achieve legal agreements (8-12)
8. Upload and standardise data (8-12)
9. OPTIONAL: develop ergonomic data discover tools for use with IHCC cohorts (7-24)
10. Standardise locally held (not uploaded) data (8-12)
11. Hold regional 'sandbox' style datathons for analysts to introduce the datasets and kick-start analyses (13).
12. Analyses underway (13-24)

**Resources**
These depends on how IHCC consider this will inform its agenda. For all options central resource for overall project management/governance, and local resource for cohort data management are assumed.

*Option 1: Use the project solely to provide evidence on current data access procedures.*

This will require data scientist, software engineer, and web-design input to DPUK. It will also require (depending on IHCC operational requirements) increased GPU and memory capacity within the Data Portal.

It is also important that analysts' time is made available to the collaborating cohorts as well as to DPUK. There has to be a scientific incentive to support this collaboration.

*Option 2: Use the project to inform and develop a new generation of ergonomic data discovery and visualisation tools.*

This will require, in addition to the resource for option 1, investment in a development programme of user engagement and software development.

*Option 3: Use the project to inform the IHCC cohort enrichment strategy*

This will require, in addition to the resource for option 1, a reserve fund from which to draw for genotyping and phenotyping. This is a strategically important option. For cohorts to see the potential for enhancement of their data is an incentive to participation

# A Pilot Study of Data Harmonization and Rare Variant Detection Among International Cohorts

**Authors:** Rongling Li, M.D., Ph.D., M.P.H., and Teri Manolio, M.D., Ph.D.

## 1. Idea summary

The purpose of this proposal is to conduct a pilot study to examine whether and how the international cohort datasets can be utilized for global genomic medicine research. The objectives of the proposal are to: 1) harmonize demographics, clinical phenotypes and genome-wide array and/or sequencing data; 2) collect high-level environmental data from cohorts at different geographic locations such as weather (temperature and humidity ranges), air quality, water quality, location efficiency (Residential, population, employment density; jobs per housing unit; employment and housing entropy, etc.), road density, and hospital accessibility; 3) impute existing genome-wide genotype array data from participating cohorts to broadest possible global reference genome; 4) sequence samples as needed from underrepresented populations such as participants in Africa Centre for Health and Population Studies (South African), Bandim Health Project (West African), Mexico City Prospective Study, and PERSIAN Cohort Study (Iran) who had biospecimens collected for genomic sequencing and agreed to share data; and 5) compare the different distributions of human knockout and clinically actionable pharmacogenomic variants, and identify rare variants associated with selected clinical phenotypes of public health importance and prevalence difference across race/ethnicity in different continents (e.g., breast cancer, prostate cancer, hypertension, T2D, HIV and/or hepatitis C infection).

## 2. Structured abstract

a. Background and rationale

It is scientifically critical and economically cost-effective to study the biological and genetic basis of diseases and improve clinical care and population health by utilizing existing as well as performing additional collection of data from international cohorts. There are 61 cohorts with a total of 30.4 million individuals in 33 countries were identified through the 1st IHCC summit (ref the white paper). Can we use data from those cohorts to conduct genomic research? If so, how? In this proposal, we aim to conduct a pilot study to test data harmonization and utilization and generate preliminary results on the distributions of known clinically actionable variants including pharmacogenetic variants and human knockout variants, as well as to assess if the same health phenotypes are associated with different rare genetic variants in different race/ethnic populations across the continents.

b. Idea

To test data utilization and possibility to collect additional data:

1. Harmonize specific data for the pilot study
2. Collect additional phenotypic and environmental data

To generate preliminary results:

1. Impute existing genome-wide array data to broadest possible global reference genome
2. Sequence (whole genome) additional samples from populations with inadequate reference genomes, as determined by available data in selected cohorts and the cost of sequencing; priority will be given to underrepresented populations with disproportionate disease burdens
3. Conduct descriptive epidemiologic studies to uncover differences in distributions of genetic variants across different race/ethnic populations

4. Assess if the same/similar phenotypes are associated with different genetic variants given different race/ethnicity and geographic locations

c. Impact

- The pilot study will provide lessons learned or best practice for future expanded international collaborations.
- The preliminary results will provide the bases for future directions.

d. Timeline

Year 1: data harmonization, additional data collection, genomic data imputation and sequencing as needed
Year 2: data cleaning and analyses
Year 3: summarization of results and results dissemination

e. Estimated funding required

Estimated budget for the pilot project is ~$18M total costs for three years. These total costs include: ~750K/year for one coordinating institution, ~$3.2M/year for 15-30 subcontracting institutions, and $6.4M for whole genome sequencing (WGS) of ~8,000 samples at ~$800/sample.

The coordinating institution will be responsible for overseeing the pilot project and coordinating all aspects to ensure that the project progresses smoothly. These include but are not limited to: 1) coordinating data standard procedure development and distributing the procedure among subcontracting institutions for implementation; 2) imputing genomic data; 3) coordinating data harmonization; 4) managing a federated data system for data sharing and analyses; and 5) coordinating network logistics such as steering committee meetings, workgroup calls, and documentation tracking.

All participating institutions, including the coordinating institution, will be responsible for: 1) participating in data standardization and harmonization of existing datasets; 2) collecting additional data as required; 3) shipping samples for WGS as needed; 4) sharing data; 5) designing analytic proposals and selecting correct methods for analyses; 6) interpreting and publishing results; and 7) disseminating lessons learned.

WGS will be conducted in underrepresented populations, such as: Africa Centre for Health and Population Studies, Bandim Health Project (BHP), Mexico City Prospective Study, and PERSIAN Cohort Study. Each study will contribute 2,000 samples for WGS.

## 3. Additional information

a. Why does this idea require multiple large cohorts?

We need to assess the difference among different race/ethnicity across different continents

b. Which cohorts would be required? [suggest that the idea submitter reach out to the cohort leaders to discuss the idea and seek their input and approval]

We will select all cohorts that have and will have genotype array and/or sequencing data available by the end of 2019. We intend to include Africa Centre for Health and Population Studies (South Africa), Bandim Health Project (Guinea-Bissau, West African), Mexico City Prospective Study, and PERSIAN Cohort Study (Iran) for whole genome genomic

sequencing, 2,000 samples from each cohort. In those cohorts, participants' biospecimens were collected and data sharing was agreed.

c. What kinds of data/sample access will be required?

We need demographic data (age, race/ethnicity, sex, geographic location including country, city and/or ranges of latitude and longitude of participants residing), genotype array and/or sequencing data, 3-5 clinical phenotypes of public health importance (considering breast cancer, prostate cancer, hypertension, T2D, HIV and/or hepatitis C infection but determined and agreed upon by collaborating cohorts)

d. What additional assays or data collection will be required?

If we will not obtain enough genotype or sequencing data from underrepresented populations, we propose to sequence participants of specific cohorts. We propose to collect environmental data through granular address information to possibly identify weather (temperature and humidity ranges), air quality, water quality, location efficiency (residential, population, employment density; jobs per housing unit; employment and housing entropy, etc.), road density, and hospital accessibility.

e. What is the data analysis plan?

Ideally, we need a centralized data descriptive and association analysis. If data sharing is a challenge, federated or meta analyses will be applied.

f. What are the technical and/or policy challenges that you anticipate will need to be addressed for this idea to be successful?

Data harmonization will be a challenge, especially if the data are in different languages. We need help from team A to overcome this challenge. We hope participating cohorts can share deidentified data. If not, we propose to share aggregated data/results.

# Addressing the Ethnicity Gap in Human Genetics Research

**Author:** Adam Butterworth, Ph.D. (University of Cambridge, UK)

## 1. Idea Summary

The vast majority of human genetic studies have included participants of European ancestry based in North America or European countries. This has led to a significant gap in our understanding of human genetics, with respect to ethnicity and geography. The consequences of this inequity are an absence of knowledge of the genetic determinants of traits and diseases in many populations, difficulties in translating clinically meaningful genetic tools (e.g., polygenic risk scores) across populations, and a substantial missed opportunity to discover novel disease-causing genes via variation that is absent (or uncommon) in European ancestry populations.

The creation of large biobanks outside of North America or Europe will help to increase the diversity of ethnic groups and countries profiled in well-powered GWAS. For example, the Japan Biobank, Kadoorie Biobank and the Tohoku Megabank are gradually building the evidence base for genetic associations in East Asia. Other studies, such as the Million Veteran Program and the *All of Us* Research Program are enrolling significant proportions of non-European ancestry participants living in the US. However, large swathes of the globe and other major (and minor) ancestral groups remain unrepresented in modern human genetics.

To help address the diversity gap in human genetics, we propose a pilot study to genetically characterise ~100,000 well-phenotype South Asian participants from IHCC cohorts, paving the way for expansion to other understudied populations.

## 2. Abstract

Background and rationale

A recent scientometric review of genome-wide association studies (Mills & Rahal, *Communications Biology*, 2019) estimated that 86% of participants are of European descent, and that >70% of all participants were recruited in the U.K., U.S. or Iceland (**Figure 1**). The rise of recent very large-scale biobanks with genetic data, such as the UK Biobank, Kaiser Permanente GERA study and deCODE, is exacerbating these inequities in ethnic composition and geographic location due to the predominant European ancestry and wide array of phenotypes resulting in many publications utilising these datasets.

The lack of ethnic and geographic diversity in genetic studies has a number of major detrimental consequences. First, while great advances have been made in our understanding of the genetic loci that influence thousands of human diseases and traits, it is rarely clear whether the same loci – and indeed same genetic architecture – apply to populations of non-European descent. Even where multi- ethnic populations have been employed, the non-European dataset has invariably been underpowered to detect between-ethnicity heterogeneity. Second, the substantial sample sizes of recent GWAS have led to the development of powerful polygenic risk scores that have predictive accuracy for common complex traits approaching that needed to provide clinical utility (Khera *et al., Nat Genet* 2018; Lee *et al., Genet Med*, 2019). However, scores derived in European ancestry participants are known to perform considerably less well in other ancestry groups (Duncan *et al., bioRxiv,* 2018; Martin *et al., bioRxiv*, 2019). Polygenic scores derived in non-European participants also perform poorly due to the much smaller sample sizes available currently in these groups. Finally, the predominance of European ancestry participants in large GWAS limits the discoveries that can be made, by preventing the opportunity of discovering associations with ethnic-specific variants or variants that are more common in other ancestry groups. Notable examples of variants identified in non-European ancestry groups are shown in **Table 1**.

Idea

To begin addressing the diversity gap in human genetics, we propose a pilot project within IHCC to genetically characterise a large (n~100,000) well-phenotyped sample of South Asian participants. Depending on the available resource, we propose two alternative scenarios for genetic characterisation:

1) Low-depth (e.g., 15X) whole-genome sequencing of all ~100,000 participants: this represents the scientifically optimal and most visionary approach, but is also the most costly. Sequencing all participants would allow extensive characterisation (e.g., ultra-rare variants, structural variants) of this ancestral group, facilitating ethnic-specific genetic discovery not afforded by reliance on array genotyping and current imputation panels;
2) Array genotyping and ethnic-specific imputation
   A South Asian-specific genotyping array and imputation panel has been developed by a consortium that has amassed ~7500 whole-genome sequences from participants of South Asian ancestry.

Impact

The resultant data could be used by the community to produce a substantial number of human genetics publications, thereby addressing the limitations of the 'Euro-centric' state of the literature currently. Successful completion of this pilot project would pave the way for similar projects in participants from other under-represented based in LMICs, e.g., Africa, South America.

Timeline

For whole genome sequencing, this would likely require the full 3 years to ship and sequence samples, generate data and call the genotypes. Additional data generation (e.g., structural variant calling, haplotype phasing etc) would need more time. For array genotyping and imputation, genotyping and calling could be completed within 18 months, with a further 6 months for data QC and imputation to both ethnicity-specific and global imputation reference panels.

Estimated funding required

To perform low depth (e.g., 15X) whole-genome sequencing on 100,000 South Asian ancestry participants would cost approximately $30-50 million. To conduct genomewide array genotyping using the South Asian array, followed by imputation, would cost approximately $3- 5 million.

## 3. Additional information

a) Why does this project require multiple large cohorts? Which cohorts would be used?

While the US National Human Genome Research Institute has recently announced $7.3 million of funding to address the lack of diversity in the human genome reference sequence, this funding will focus on generation of high quality sequence data in small numbers of participants from diverse genomic backgrounds. While this will ultimately enable higher quality imputation in non-European ancestry participants, the limited sample size involved will not address the limited proportion of non- European participants in human genetic studies.

We propose to start addressing the diversity gap by aiming to sequence (or genotype and impute) 100,000 participants from South Asian ancestry cohorts. The rationale for focusing on South Asian participants is:

- Availability of potential cohorts in IHCC: there are multiple South Asian cohorts in IHCC that have a large sample size, available DNA and consent for sequencing, linkage to health phenotypes and consent to recontact participants for follow-up studies;
- Under-representation in GWAS: although ~10% of participants in GWAS are of Asian ancestry, the vast majority of these are East Asian (predominantly Chinese or Japanese). Recently established in China and Japan will exacerbate this inequity.
- Lack of national funding: numerous countries around the world are now investing significantly in national genomics strategies, including the U.K. (100,000 Genomes Project), France

(Médecine Génomique 2025), Qatar (Qatar Biobank), Denmark (Genome Denmark), Saudia Arabia (Saudi Human Genome Program). However, within the LMICs in South Asia (e.g., Bangladesh, India, Pakistan), no nationally funded genomics projects exist.

- History of intermarriage: populations that have a history of familial intermarriage for social and cultural reasons, such as South Asian or Middle Eastern populations, have higher rates of autozygosity than outbred Europeans, leading to higher numbers of "human knockouts", participants who are homozygous for loss-of-function alleles (Saleheen *et al., Nature*, 2017; Narasimhan *et al., Science*, 2016). Sequencing studies of these populations are far more efficient for identifying these highly informative participants, who provide insights into the effects of completely knocking out proteins.

b) What kinds of data/sample access would be required?

Access to DNA samples would be required and permission to send these samples to international sequencing hubs (e.g., Wellcome Trust Sanger Institute, Broad Institute etc) would be required to make the project feasible. To generate the association data to begin addressing the ethnicity gap, access would then also be required to individual-participant data on the broadest possible array of phenotypes. Initially, we anticipate analyses could be conducted within-cohort, limiting the need to share full raw data across international boundaries, but best working practices would be evaluated (with input from Teams A and C) as part of this pilot project to inform future such IHCC projects.

c) What additional assays or data collection would be required?

At the IHCC kick-off meeting in 2018, there was considerable discussion about the value of array genotyping vs whole-exome sequencing vs whole-genome sequencing in IHCC cohorts, particularly those from LMICs without national genomics strategies. These discussions concluded that "*Most valuable would likely be whole genome sequencing with sharing of WGS data files, while WES (identifying new variants in known GWAS genes) did not seem a high priority….interest among non- European cohorts was greater for array genotyping supplemented by subgroup WGS.*" Hence we propose to enrich cohorts within IHCC of substantial non-European ancestry with WGS data.

d) What is the data analysis plan?

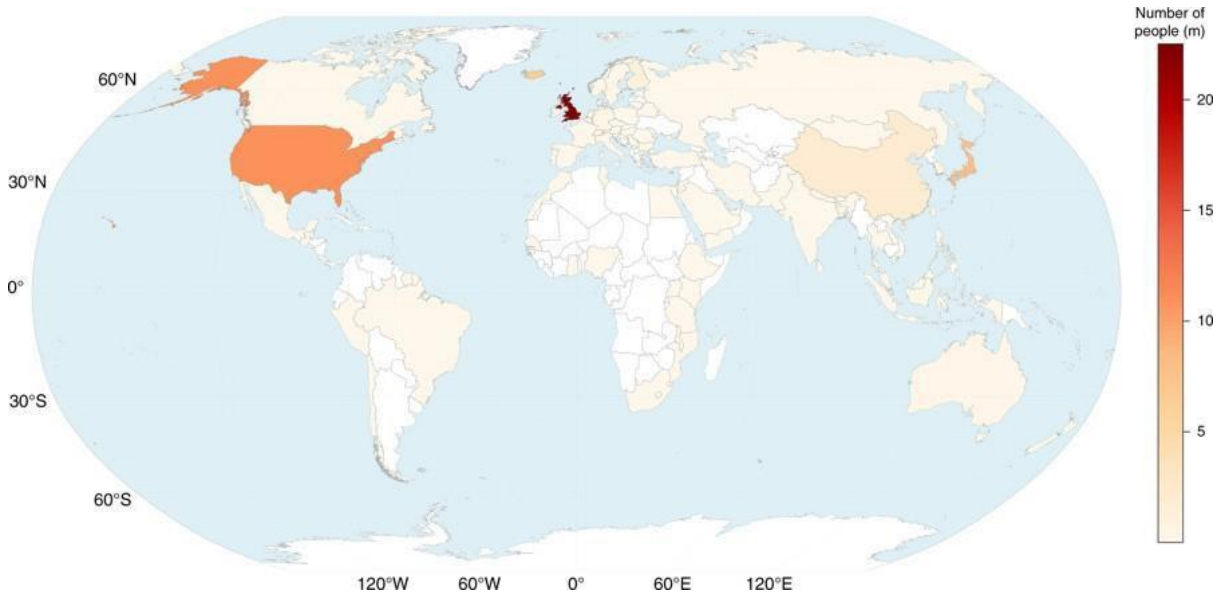
Resulting data would be fed into external initiatives so that the information generated can be maximally used by the scientific community. For example, sequencing data should be make publicly available (e.g., via dbGAP or EGA), data should be fed into the Genome Aggregation Database (gnomAD) to facilitate estimates of allele frequencies in these understudied populations, and phased data should flow into ethnic-specific and global reference panels (e.g., the Haplotype Reference Consortium).

e) What are the technical and/or policy challenges?

Technical challenges would include infrastructure to call, store and analyse such large-scale genomic data and to make the data widely available to the scientific community. However, these challenges are being addressed by other initiatives (e.g., UK Biobank, *All of Us*, etc.) so will not be insurmountable. Other challenges involved with working in LMICs would relate to appropriate consent to use the DNA for the requested purposes and to send the DNA overseas to a major sequencing hub, as well as to appropriately involve local investigators and the local community (e.g., training local scientists to work with these data).

Tables and Figures

**Figure 1. Global map highlighting the lack of geographic diversity in GWAS participants**




(reproduced from Mills & Rahal, *Communications Biology*, 2019)

**Table 1. Notable examples of pivotal genetic discoveries made in non-European ancestry groups**

| **Gene** | **Ethnic group** | **Trait/disease** |
|---|---|---|
| *APOC3* | Amish | Lipids/heart disease |
| *PCSK9* | African-American | Lipids/heart disease |
| *MYBPC3* | Indian | Cardiomyopathy |
| *PLA2G7* | East Asian | Lp-PLA2 |
| *TBC1D4* | Greenlanders | Type 2 diabetes |
| *ASGR1* | Icelanders | Lipids/heart disease |

# Applications of Mitochondrial Haplogroups in Understanding Disease Risk Across Different Ethnic Groups

**Lead Investigator:** Xiao Chang, Ph.D.

**Primary Cohort Site:** Center for Applied Genomics Cohort at Children's Hospital of Philadelphia (CAG Cohort @CHOP, laboratory of Dr. Hakon Hakonarson)

## 1. Idea Summary

Identification of mitochondrial DNA (mtDNA) variations as risk factors for chronic complex diseases represents a critical outcome for genomic medicine in that it provides a forum for (individualized) screening and risk prediction, as well as tools for exploring functional/etiological disease mechanisms. Pairing the study of mtDNA variation with multiple traits represents and interesting opportunity, but requires large sample-set for requisite power. At the Center for Applied Genomics (CAG) of the Children's Hospital of Philadelphia (CHOP), we have identified multiple novel mtDNA variants in the most prevalent diseases of childhood, such as autism, ADHD and neuroblastoma. Building upon these findings, we aim to further examine the impact of mtDNA in adults and multiple ethnicities across several large cohorts, which have direct implications for drug development, prescribing, and population screening.

## 2. Structured Abstract

Background and rationale:
Mitochondria are cellular organelles participating in bioenergetic metabolism and producing adenosine triphosphate (ATP) through oxidative phosphorylation (OXPHOS). Dysregulation in OXPHOS has been reported in multiple genetic diseases such as diabetes, cancers and neurodegenerative diseases, indicating a role of mtDNA variation in multiple genetic diseases. Although Intensive efforts have been made in discovering predisposition genes from nuclear DNA (nDNA) through genome-wide association studies (GWAS) and whole genome/exome sequencing studies, little attention has been paid to mtDNA variations, which could contribute to the 'missing heritability' of genetic diseases. Our center is currently investigating the association between mitochondrial variations and multiple pediatric diseases. We want to contrast our findings with the corresponding traits seen in adults and other populations. As such, we would like to employ genotype and sequencing data from IHCC to validate our most significant signals and to extend some of extend our efforts into adult populations and new disease areas. We have several unpublished hypotheses that we would like the IHCC datasets to help inform.

Idea:
We aim to further examine the impact of mtDNA variations in chronic complex diseases across several large cohorts. The human mitochondrial genome sequentially accumulates genetic variants through maternal inheritance. As a result of ancient migrations of human populations, mtDNA variants have segregated and clustered into regional groups of related mtDNA haplotypes, called haplogroups. These haplogroups vary significantly in frequencies across continents, and exhibit diverse metabolic capacities. We will focus on studying the association between haplogroups and genetic diseases, and identify the precise causal variants across multiple diseases.

Impact:
We will build upon our extensive experience in genetics, bioinformatics and mitochondrial diseases to devise and validate statistical strategies to identify causal variations/haplogroups from mtDNA. These innovative methods will have broad application in many mitochondria-associated complex diseases. We will also conduct functional validation experiments to evaluate the effect of mtDNA changes on gene expressions and mitochondrial pathways by using animal models or cell lines. Discoveries made from our data analytics may lead to new methods for the prevention and treatment of diseases.

Timeline:
The proposal analysis will be completed in three years. We anticipate the genotyping array data analysis to be completed by year 1. Sequencing data analysis will be completed by year 2. Functional follow-up will be completed by year 3.

Estimated funding required:
$450,000 for all 3 years. Pending availability of funds, cost sharing could be considered for the project.

### 3. Additional Information

Why does this idea require multiple large cohorts? Which cohorts would be required?
The importance of this idea lies in its large population-based sample of participants sourced from multiple communities, rather than an exclusively disease-specific group, which could potentially be subject to selection bias. The increased sample size can also give us greater power to detect disease-causal variants. While this project can go forward with any number of cohorts. In addition to our CHOP cohort, we find the following cohorts to be of best match and would like to require access to them:

23andme, US Biobank Japan
Cancer Prevention Study-II
(CPS-II) China Kadoorie
Biobank
China PEACE
Estonian Genome Project
Genomics England / 100,000 Genomes
Project Kaiser Permanente Research
Program
Korea Biobank Project
Korean Genome and Epidemiology
Study Korean Cancer Prevention Study-
II (KCPS-II) LifeGene (and sister cohort,
EpiHealth), Sweden Million Veteran
Program, US
MyCode Community Health Initiative,
US Newfoundland and Labrador
Genome Project Northern Sweden
Health and Disease Study Norwegian
Mother and Child Cohort Study
Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial
(PLCO, NCI) UK Biobank
UK Blood Donor Cohorts
UK Collaborative Trial of Ovarian Cancer
Screening Taiwan Biobank
23andme, us Biobank Japan

Required data/sample access:
Genetic data (sequence, genotypes) and clinical phenotypes.

Additional assays or data collection required:
We do not anticipate need for new data collection or specific assays to be required.

What is the data analysis plan?

Study Objectives

The aim of this study is to identify the effects of mtDNA variations to the risk of complex diseases. The primary objective is assessing whether mitochondrial haplogroups among these study populations are associated with specific diseases. Secondary objectives include exploring the precise causal variants from disease-associated haplogroups, and investigating the relationship between detected mtDNA variants/haplogroups with known disease-associated nDNA variants across multiple different ethnic groups.

Study Design

Population/cohort demographics

Cases and controls will be determined based on available clinical information and records from each site. Variables such as ages and gender will be taken into account in the statistical analysis. We will use Multi-Dimensional Scaling (MDS), as implemented in the PLINK software, for inferring population structure in the genotyping/sequencing data set. Reference data sets will be downloaded from 1000 Genomes. The identity- by-descent (IBD) scores will be calculated to remove cryptic relatedness between samples. One individual in the pairs of subjects was excluded with IBD greater than 0.25.Data Analysis:

Genotyping array data and sequencing (whole genome sequencing or whole exome sequencing) data will be analyzed separately.

For genotype data (genotyping array), we will perform a genotype imputation pipeline using the 1000 genome reference data to predict the missing genotypes of mitochondrial SNPs, and use HaploGrep tool to deduce haplogroups by identifying the best match of genotyped and imputed variants to the variants in the mtDNAs encompassed within the global phylogeny of complete mtDNA sequences. Logistic association test will be performed in R/PLINK to evaluate the association between haplogroups and diseases. Gender and the first ten principal components will be included as covariates.

For sequence data (whole genome/exome sequencing), we will also detect mtDNA heteroplasmy and evaluate its role in genetic diseases. Sequencing reads passing the quality filter will be aligned to the human reference genome (GRCh37-derived alignment set used in 1000 Genomes Project) with Burrows-Wheeler Aligner.

Sequencing reads mapped to the mitochondrial genome will be included in the downstream analysis. For homoplasmic as well as heteroplasmic variant detection, reads with a mapping quality score <20 (Phred score) and a read length <25 will be removed. BAM reads marked as duplicates will also be filtered within this step.

The GATK BAQ implementation will be adapted to work with the circular nature of mitochondrial genomes. For heteroplasmy detection, several filters and methods will be applied: first, mitochondrial hotspots around 309 and 315 as well as 3107 according to the rCRS will be excluded. Sites showing coverage <10 bases per strand will be filtered. For all remaining sites showing (i) a VAF of ≥ 1% (strand independent) and (ii) an allele coverage of three bases per strand, an ML model will be applied. The ML model takes sequencing errors per base into account and will be applied to each strand. All sites with a log likelihood ratio (LLR) of ≥5 will be tagged as heteroplasmic sites. Since strands are analyzed independently, all heteroplasmic sites will be filtered with a strand bias score <1. The Wilson and the Agresti-Coull confidence interval will be calculated for heteroplasmic variants. Finally, haplogroups will be deduced by haplogrep based on the detected mtDNA variants. Similarly, Logistic association test will be performed in R/PLINK to evaluate the association between haplogroups/SNVs (single nucleotide variants) and diseases. Gender and the first ten principal components will be included as covariates.

Epistasis is the phenomenon where the phenotypic effect of variation at one genetic locus depends on variation at other loci. We will also explore the potential epistasis effect between mtDNA variants and detected nDNA variants in diseases. All pairwise interactions among known genetic loci and detected mtDNA variations will be assessed in each disease-specific dataset using two different approaches

FastEpistasis and BOOST, which are implemented in PLNK.

We will also generate RNAseq data to validate the functional impact of mtDNA variations in complex diseases. The association between risk variants/haplogroups and expression levels of genes encoded by mitochondrial genome will be investigated. Genomic Short-read Nucleotide Alignment Program will be used to map the reads to the reference human genome (hg19). Common single-nucleotide polymorphisms (SNPs) recorded in dbSNP will be taken into account to improve the alignment accuracy. Both fragments per kilobase of transcript per million mapped reads (FPKM)-based and count-based algorithms will be used to evaluate the gene expression level. The FPKM-based algorithm is implemented in Cufflinks and count-based algorithms is implemented in R packages featurecounts.

Technical and/or policy challenges that will need to be addressed:
Data-sharing across cohorts will be challenging but can be readily overcome.

# Application of Polygenic Risk Scores (PRS) for Improved Health Outcomes in Alzheimer Disease, Cardiovascular Disease and Across Multiple Neuropsychiatric-Phenotypes

**Lead Investigator:** Patrick Sleiman, Ph.D., Associate Director of the Center for Applied Genomics (CAG)

**Primary Cohort Site:** Center for Applied Genomics Cohort at Children's Hospital of Philadelphia (CAG Cohort @CHOP, laboratory of Dr. Hakonarson)

## 1. Idea Summary

Polygenic risk scores (PRS) have been shown to be highly predictive of complex phenotypes with underlying polygenic inheritance involving many common genetic variants of small effect as opposed to rare monogenic mutations. By applying a trans-ancestry approach to the genetic relationship between complex phenotypes, we have a major opportunity to further delineate 1) the heritability of relevant disorders, 2) the extent to which individuals of different ancestries differ/share genetic risk scores, 3) underlying biological mechanisms, and 4) diagnostic/screening that may be used to identify predispositions earlier (and therefore targeted therapeutically). We will apply the approach to several disease areas where 1) a relatively large number of genotyped participants already exist across the IHCC, 2) a relatively large number of well-defined phenotypes already exist across the IHCC, and 3) the literature has a relatively consistent well-defined and validated set of biomarkers. Alzheimer's disease, cardio-, and (certain) neuropsychiatric-phenotypes are immediate candidates from this perspective. Several approaches to PRS are considered, with the pruning and thresholding (P+T) and LDPred algorithm methods arguably the most powerful. We propose to generate PRS using both approaches with varying tuning parameters. As PRS are highly impacted by changes in linkage disequilibrium we will develop scores separately for Asian, African, European-Hispanic, and European Non-Hispanic ancestry populations. Both approaches require a training set from which we derive the GWAS summary statistics from which the models are developed and a separate validation set to determine the predictive accuracy. While certain technical and data-sharing challenges must be negotiated, we anticipate that these projects would deliver a high return on investment insofar as they would represent important contributions to the literature on risk and etiology with clear implications for clinical practice.

## 2. Structured Abstract

Background and rationale:
In a recent publication, Kathiresan et al.(1) generated genome-wide risk scores for five common diseases, coronary artery disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease and breast cancer. The scores were constructed from over 6 million variants are were shown to highly predictive of disease risk, the area under the curve (AUC) for CAD was 0.81 which translates to 8% of the population with greater than threefold increased risk of CAD. The authors compared three approaches to generate PRS, the first was using GWAS significant variants alone, the second was the widely applied pruning and thresholding (P+T) method and the LDPred algorithm(2) a Bayesian approach that accounts for linkage disequilibrium (LD) to increase predictive accuracy. LDPred yielded higher AUC values for all diseases except breast cancer where the P+T approach was marginally more predictive.

Idea:
We aim to take a similar approach to Alzheimer's disease, cardio-, and neuropsychiatric-phenotypes. We propose to generate PRS using both the P+T approach and LDPred with varying tuning parameters. As PRS are highly impacted by changes in LD we will develop scores separately for our African American

and European ancestry populations. Both approaches require a training set from which we derive the GWAS summary statistics from which the models are developed and a separate validation set to determine the predictive accuracy. Case-control datasets will be defined by electronic algorithm, with sensitivity to the depth and quality of phenotype data available at respective sites, with a premium on structured data derived from electronic health records and medication history. We will use CHOP and other IHCC cohorts for the training sets, with other geo-match IHCC sites for validation.

Impact:
With increasingly larger sample sizes, the GWAS approach has been increasingly able to capture larger proportions of individual variation, though of course this is phenotype- and ancestry-dependent. The main utility of the PRS approach is in identifying individuals with particularly high (or low) genetic risk, where the combination of small effects, can approach that of individuals with monogenic disease (1).

Timeline:
Year 1:
Phenotype definitions leveraging electronic algorithms across sites:
- Alzheimer Disease (AD)
- Cardio-phenotypes

Year 2:
Phenotype definitions leveraging electronic algorithms across sites:
- Neuro- phenotypes

Complete AD analysis, submit for publication

Year 3:
Complete PRSs for:
- Cardio- phenotypes
- Neuro- phenotypes

Complete PRSs, submit for publication

Estimated funding required:
$300,000 - $400,000 in total funding. Pending availability of funds, cost-sharing can be considered.

**3. Additional Information**

Why does this idea require multiple large cohorts? Which cohorts would be required?
This project can go forward with any number of cohorts. In addition to our CHOP cohort, we find the following cohorts to be of best match and would like to require access to them:

23andme, US
China Kadoorie Biobank
Estonian Genome Project
Genomics England / 100,000 Genomes Project
Kaiser Permanente Research Program
Korea Biobank Project
Korean Genome and Epidemiology Study
LifeGene (and sister cohort, EpiHealth), Sweden
Million Veteran Program, US
MyCode Community Health Initiative, US
Norwegian Mother and Child Cohort Study
UK Biobank
UK Blood Donor Cohorts
UK Collaborative Trial of Ovarian Cancer Screening

Required data/sample access:
Genetic data
- Genotypes
- Exome sequence
- Genome sequence
- Targeted panel

Phenotype data
- EHR-derived or standardized measures preferable
- Medication history

Additional assays or data collection required:
We do not anticipate need for additional assays.

What is the data analysis plan?
We propose to generate PRSs using both the method and LDPred algorithm with varying tuning parameters.
For the P+T approach we will carry out informed LD pruning, which preferentially prunes the less significant marker, yielding much more accurate predictions than pruning random markers followed by p value thresholding, i.e., including only markers that achieve a given significance threshold. We will carry out LD pruning with an r2 threshold 0.2 and subsequently applying p value thresholding, where the p value threshold is optimized over a grid with respect to prediction accuracy in the validation data.

LDpred has a several advantages over P+T, first we are not required to LD prune the dataset. LD pruning discards informative markers and thereby limits the overall heritability explained by the markers. Second, LDpred accounts for the effects of linked markers, which can otherwise lead to biased estimates. These limitations hinder P+T regardless of the LD pruning and p value thresholds used. By using a point-normal mixture prior for the marker effects, LDpred can be applied to traits and diseases with a wide range of genetic architectures. Unlike P+T, LDpred has the desirable property that its prediction accuracy converges to the heritability explained by the SNPs as sample size grows.

LDpred estimates posterior mean causal effect sizes from GWAS summary statistics by assuming a prior for the genetic architecture and LD information from a reference panel. As suggested by the developers (2) we will use the validation cohort as the LD reference panel. The only remaining variable is the tuning parameter which models the proportion of variants assumed to be causal, we will apply a sliding scale from 0.001 to 1, i.e. assuming at the lower end that 0.1% of variants are causal and increasing to all variants. Final choice of tuning parameter will be based on the highest prediction accuracy in the validation data. Generating individual risk scores: Having trained the models and optimized the tuning parameters, individual risk scores can be generated to identify individuals at increased risk of OUD.

Technical and/or policy challenges that will need to be addressed:
Phenotypes: Defining case and control datasets across cohorts and continents presents challenges. While a centralized phenotype repository would provide the ideal platform from which to build, we assume this capacity would not available in the immediate term. As such, a federated approach where collaborators implement collaboratively-developed algorithms at a local site provides the most pragmatic approach to delivering a quick win. [Note: we would fully support a parallel effort where relevant datasets are available to the IHCC for standardization pilots or similar efforts].

Electronic algorithms will be the primary driver for case/control data selection, using a combination of diagnostic codes and prescription history. We have also expertise in aggregating unstructured textual data and the search or NLP tools to mine data for phenotypic information, which can be shared with collaborators to unstructured free-text (clinical letters, progress notes, reports, discharge summaries, etc.).

Linkage Disequilibrium: Accounting for linkage disequilibrium (LD) across populations is a complicating factor for trans-ancestry analyses, although recent advances have improved the methods by which these analyses can be conducted. For each phenotype, major ancestry groups will be analyzed separately (though a meta-analysis of each may be informative).

Interpretation:
It is important to acknowledge that PRS-defined risk will likely not reflect a single underlying mechanism, but rather a combination of pathways, which will be difficult to tease apart. While this represents an undoubted challenge, it is also likely that scores will have utility regardless of the bio-mechanism, as is often the case with screening for coronary artery disease, and many types of cancers.
Communication: For a methodology in its relative infancy, we need to be especially careful in promulgation and commination of risk scores to patients and providers, which must be carefully weighed against their actionability. This is particularly evident with AD, where return-of-results is not routine, and the research-ethical “principle of caution” recommended (3, 4).

**Bibliography**


1. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, Kathiresan S. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018;50(9):1219-24. Epub 2018/08/15. doi: 10.1038/s41588-018-0183-z. PubMed PMID: 30104762; PMCID: PMC6128408.

2. Vilhjalmsson BJ, Yang J, Finucane HK, Gusev A, Lindstrom S, Ripke S, Genovese G, Loh PR, Bhatia G, Do R, Hayeck T, Won HH, Kathiresan S, Pato M, Pato C, Tamimi R, Stahl E, Zaitlen N, Pasaniuc B, Belbin G, Kenny EE, Schierup MH, De Jager P, Patsopoulos NA, McCarroll S, Daly M, Purcell S, Chasman D, Neale B, Goddard M, Visscher PM, Kraft P, Patterson N, Price AL. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am J Hum Genet. 2015;97(4):576-92. Epub 2015/10/03. doi: 10.1016/j.ajhg.2015.09.001. PubMed PMID: 26430803; PMCID: PMC4596916.

3. Erdmann P, Langanke M. The Ambivalence of Early Diagnosis - Returning Results in Current Alzheimer Research. Curr Alzheimer Res. 2018;15(1):28-37. Epub 2017/09/12. doi: 10.2174/1567205014666170908101237. PubMed PMID: 28891443.

4. Frieser MJ, Wilson S, Vrieze S. Behavioral impact of return of genetic test results for complex disease: Systematic review and meta-analysis. Health Psychol. 2018;37(12):1134-44. Epub 2018/10/12. doi: 10.1037/hea0000683. PubMed PMID: 30307272; PMCID: PMC6263735.

# Atopic diseases, early Life Exposure Risk factors, and Genetics in the Young (ALERGY) study

**Authors:** Elizabeth Jensen, M.P.H., Ph.D.; Lindsay Reynolds, Ph.D.; Mara Vitolins, Dr.P.H., RDN; Michael O'Shea, M.D., M.P.H.

## 1. Idea Summary

Evidence suggests that the incidence of atopic disease in children is increasing. One hypothesis that has been advanced for explanation of this increase is the potential that potentially immune-modifying perinatal and early life exposures, in genetically susceptible children, increases risk for atopic diseases. Assessments of gene-environment interaction necessitate very large cohorts, a limitation of studies conducted to date. The main objective of this pilot project is to characterize the descriptive epidemiology of atopic diseases (atopic dermatitis, allergic rhinitis, asthma, and food allergies) in children, evaluating possible etiologic factors contributing to these diseases. To achieve this objective, we propose to develop a common data model for data harmonization, specifically leveraging data across multiple cohorts, including the Japan Environment and Children's Study, Korean Children's Environmental Health Study, Norwegian Mother and Child Cohort Study (MoBa), Danish National Birth Cohort, and the Environmental Influences of on Child Health Outcomes cohorts. In leveraging multiple pediatric cohorts, this pilot study will be well-positioned to conduct a future study examining gene-environment interaction in relation to development of atopic disease in children.

Evidence from candidate gene studies with small sample sizes suggest genetic variants within human leukocyte antigen (HLA) molecules, encoded within the major histocompatibility complex (MHC) located on chromosome 6, and other candidate genes, may alter atopic disease risk. In phase II of the proposed study (after the pilot), we plan to examine common variants in these genes as potential effect modifiers of environmental risk factors associated with the development of atopic disease. With a sufficiently large sample size, we will perform a genome-wide association study to unbiasedly identify potential genomic susceptibility loci that may modify the risks associated with atopic disease.

## 2. Structured Abstract

Background/rationale: In the United States, population-based epidemiologic data indicate that the prevalence of atopic conditions vary by age, race/ethnicity, sex, and socio-economic status and that the prevalence may be increasing among boys and girls, children of non-Hispanic white and non-Hispanic black race, and Hispanic children. However, the extent to which differences in health-related knowledge and behaviors, access to medical care, distributions of risk factors, or geographic exposures contribute to observed differences in the incidence and prevalence of atopic diseases is still unknown. A prevailing hypothesis is that perinatal and early life exposures have the potential to dysregulate immune development and that in the setting of genetic susceptibility, increase risk of developing atopic disease. While numerous studies have sought to examine these associations, including evaluation of possible gene-environment interaction, small sample sizes have limited the interpretation of these studies.

Idea: We will apply robust epidemiologic methods and leverage existing pediatric cohorts to describe and contrast the descriptive epidemiology of atopic disease across pediatric cohorts. Perinatal and early life factors will be examined in relation to development of disease. A common data model will be used to harmonize data across cohorts. This pilot will pave the way to a second phase of the study, for which biobanked samples will be genotyped and/or existing genotype data will be leveraged for a gene-environment interaction assessment. We propose to:

(1) Evaluate the incidence, prevalence, and predictors of atopic disease among and between children in the the Japan Environment and Children's Study, Korean Children's Environmental Health Study,

Norwegian Mother and Child Cohort Study, Danish National Birth Cohort, and the Environmental Influences of on Child Health Outcomes cohorts.

(2) Estimate the association between risk factors, including demographic, health-related, geographic, perinatal and early life factors, of childhood atopy

(3) Assess modification of the effects of these risk factors by genotype (phase II)

Impact: This project seeks to improve our understanding of the epidemiology of atopic disease in children. This research proposes to use traditional epidemiologic methods to characterize the incidence, prevalence and etiologic factors contributing to childhood atopy, an important public health problem of growing concern. The anticipated results will provide novel information with which to enhance our current understanding of the epidemiology of atopic disease among children, including identification of possible modifiable factors for disease prevention. The inclusion of susceptibility genotype may yield opportunity to identify novel risk factors for disease, specifically factors only indicative of increased risk in those with underlying genetic susceptibility. Conversely, genetic susceptibility may only become apparent in the setting of certain environmental exposures. These susceptibility factors could help inform possible targets for pharmaceuticals designed to mitigate disease.

Timeline: The proposed project will span 36 months and will include the following components: 1) Development of the common data model for data harmonization, (2) Construction of analysis files and creation of derived variables (3) Primary statistical analysis; (4) Preparation, revision, and submission of abstracts and manuscripts; (5) Conference presentations (AAAAI meeting, Pediatric Academic Societies Meeting)

Estimated funding required: ~500K

### 3. Additional Information

Investigators: Members of this research team are uniquely qualified to conduct this research. We are a multi-disciplinary team of investigators with expertise in epidemiology, genetics, nutrition, and pediatrics. Dr. Jensen is a reproductive, perinatal, and pediatric epidemiologist and has extensively reported on early life exposures in relation to development of eosinophilic esophagitis, an increasingly common, allergic-mediated disease. Her study of gene-environment interaction in development of EoE identified a possible protective association for breastfeeding, but only in the presence of genetic susceptibility in the CAPN14 gene. Dr. Jensen has also contributed to development of an international, data harmonization study comparing pediatric diabetes in the U.S. SEARCH for Diabetes study and the India-based registry of youth onset diabetes. This project utilized the Observational Medical Outcomes Partnership (OMOP) model for data harmonization. Dr. Jensen has an on-going collaboration leveraging registry-based data in Denmark and has used the Mother Child Cohort study (MoBA) in Norway to evaluate anthropometric outcomes. Dr. Vitolins is a Professor with training in preventive care and wellness and is a registered and licensed dietitian/nutritionist who is co-investigator on the WHI trial (IHCC partner). She has training in dietary assessment and has overseen the dietary data collection for NIH funded studies including the Rural Undernutrition study, the Insulin Resistance Atherosclerosis study and the Hemochromotosis and Iron Overload Screening Study. Specific to this pilot study, these experiences have enhanced her knowledge of the strengths and limitations of dietary data and food frequency data analysis.
Dr. Reynolds is a genetic epidemiologist whose research aims to better understand the underlying pathways which link environmental exposures to the development of disease. Dr. Reynolds has experience integrating multiple layers of genomic data collected by large cohort studies to characterize genomic, transcriptomic, and epigenomic features of aging, age-related disease and disease risk factors.

We will utilize data collected from population-based maternal and pediatric cohorts in Japan, Korea, Norway, Denmark and the United States. Demographic characteristics, health and functional status, perinatal factors, early childhood, middle childhood and adolescence, parental health, and neighborhood/community

characteristics about the child will be harmonized. These data, which will include responses to questionnaire items concerning atopic dermatitis, allergic rhinitis, asthma, and food allergies, thus providing a unique opportunity to glean new insights into atopic diseases in children. To address the main objective of this proposed project, the research team will harmonize data from existing cohorts to characterize the descriptive epidemiology of atopic disease overall and across cohorts, and evaluate risk factors for disease, including demographic and environmental factors, and perinatal and early life exposures in particular.

Factors to be harmonized across cohorts include:

*Maternal factors: demographic factors, perinatal events including maternal infection, measures of adiposity, health history including history of atopic disease, antibiotic use and use of PPIs in pregnancy, smoking and tobacco use, diet, alcohol use*

*Paternal factors: demographic factors, measures of adiposity, health history including history of atopic disease, smoking and tobacco use, diet, alcohol use*

*Childhood factors: demographic factors, early life exposures including NICU admission, cesarean delivery, gestational age at delivery, weight for gestational age, breastfeeding and diet, anthropometrics and growth, antibiotics and use of PPIs in infancy*

*Additional factors: year of birth (for evaluation of possible cohort effects across time), environmental factors including geographic location, rural versus urban residence, climate, season of birth*

*Biospecimens – The cohorts selected either have existing biospecimens from which samples can be genotyped or have the potential for obtaining samples for genotyping. As part of the study pilot, we will develop the protocol for obtaining and genotyping samples from each of the cohorts, and harmonizing genetic data across cohorts.*

The collaboration and the resulting manuscripts will be valuable steps towards advancing research, expertise, and experience in childhood development of atopic disease.

Statistical Analysis:

Statistical analysis will be performed at Wake Forest School of Medicine. We will first characterize the descriptive epidemiology of atopic disease conditions in the respective cohorts, comparing across cohorts and standardizing against the underlying population structure to estimate differences by country. For analyses examining possible etiologic factors, harmonized data will be used to first generate descriptive statistics to summarize the distributions of possible exposures (e.g. infant antibiotic use, weeks of gestation at delivery, NICU admission, delivery mode, duration of breastfeeding, furred pet ownership in infancy) and each predefined candidate covariate of interest. To test the hypothesis that anti- and postnatal factors are associated with atopic disease, we will use generalized linear models to estimate the risk of atopy for each exposure. Candidate covariates for inclusion will be those factors that are 1) associated with the early life factor of interest in the underlying source population at risk for atopy, and 2) associated with atopy within the reference level of the exposure. Possible covariates include maternal marital status, age, education, and parity. Ante- and intrapartum factors may be share the same pathway(s) to disease development, and primary analyses, while sufficient for calculating total effect may not necessarily disentangle direct from indirect effects. As a first step, we will assess for interaction between the exposure and the possible mediating factor, through introduction of an interaction term in the model. If interaction is not present, we can include the possible intermediating factor in the model to assess for direct effects. If interaction is evident, than we will apply Vanderwheel's effect decomposition method for mediation analyses to estimate the direct effect of each exposure on atopy. We will also consider whether an association between postnatal antibiotic use and atopy may be an artifact of confounding by indication, and will use this same mediation analyses approach to assess whether antibiotic use is an intermediate between perinatal factors (i.e. preterm delivery) and

development of atopy. The proposed analyses will be performed using SAS version 9.2 (SAS Institute, Inc., Cary, NC).

**Technical challenges:** We recognize the challenge of harmonizing data across cohorts. The granularity of exposure, covariate and outcome data may be reduced through the harmonization process. However, despite this challenge, the integration of these data with genotype has the potential to provide the largest study to date of gene-environment interaction in the setting of atopic disease and advance understanding of the increasing incidence of these conditions.

# Clonal Hematopoiesis and Global Aging

**Author:** Francine Grodstein, Sc.D.

## 1. Idea summary

Clonal hematopoiesis (CH), resulting from an expansion of cells derived from a single hematopoietic stem cell, is prevalent in 10-15% of the population over age 70 years, and is associated with cancer, cardiovascular disease and overall mortality in both mid-life and older ages. We propose to investigate somatic mutations in candidate driver genes thought be responsible for CH, and their role in chronic disease. These can be assessed by relatively deep sequencing (800-1000x) of a specific panel of genes (ranging from 20-70 genes, depending on the panel). Leveraging the unique diversity of cohorts in the IHCC, we propose to investigate differences in CH-related mutations across populations that span a breadth of racial/ethnic, geographic, and socioeconomic representation. Assays are simple to do using stored DNA samples, and participating cohorts would be selected to represent a large span of ages, and a diversity of racial/ethnic groups across many countries.

## 2. Structured abstract

Background and rationale: Hematopoietic stem cells provide blood cells throughout life. Clonal hematopoiesis (CH), which can result from somatic mutations driving clonal expansion of blood cells, is a hallmark of hematologic cancers. However, it has recently become appreciated that CH becomes increasingly common with aging, with an estimated 10-15% prevalence in those over age 70 years, and that hematologic cancers are increased but not determined by CH prevalence. Indeed, CH mutations have been related to hematologic cancers, but also cardiovascular diseases, and overall mortality in initial epidemiologic research over the last 5 years. CH appears to drive inflammation and immune function – risk factors across many chronic diseases - however, little is known regarding factors related to CH-related mutations. Thus, a large investigation of the prevalence, and related factors, as well the range of health consequences associated with mutations is needed across a large variety of ages, and racial and ethnic groups.

Idea: We propose to use a CH panel assay to measure somatic mutations. CH is assessed by relatively deep sequencing of a limited number of selected genes (e.g., 800-1000x, 20-70 genes). By leveraging stored blood/DNA samples available across cohorts in the IHCC, across multiple countries and continents, we can investigate differences in CH-related mutations by country, race/ethnicity, sex, and age. We can also examine risk factors related to CH mutations, such as obesity, hypertension, and cardiometabolic diseases. Finally, as follow-up after blood collections continues in cohorts, we can begin to compare CH in incident cases versus controls selected on the basis of specific chronic diseases (i.e., hematologic cancers, non-hematologic cancers, cardiovascular diseases, cerebrovascular diseases).

Impact: Better understanding CH-related mutations in the population could reveal novel mechanisms underlying health in aging, and lead to identification of interventions that could impact health worldwide, especially as life expectancy increases throughout many countries.

Timeline: The work involved in this proposed idea would include: 1) CH panel in archived blood samples within each participating cohort; 2) detailed analysis plans to be carried out within each cohort, including risk factor association analyses and case-control analyses for specific diseases, where possible; 3) meta-analysis of results across cohorts. We anticipate the CH assay would be completed in year 1, with concurrent development of analytic plans. Analysis of data would be conducted in year 2.

Estimated funding required: CH assay of peripheral blood cells would be required, at a cost of approximately $100-200 per sample. Funding required for individual cohort needs would vary

depending on infrastructure and cohort size, but would need to cover accessing and shipping blood samples, data management to incorporate CH results with other cohort data, and analytic programming to conduct analyses. Additional funding may be required to cover the time of investigators coordinating the efforts within individual cohorts as well as investigator time to carry out this consortium project.

# Genetic and Non-genetic Risk Factors for Uncommon Cardiometabolic Conditions

**Author:** Adam Butterworth, Ph.D. (University of Cambridge, UK)

## 1. Idea Summary

Large consortia of prospective cohort studies have characterized in great detail the risk factors associated with common complex cardiovascular outcomes, such as coronary artery disease (CAD) and ischemic stroke. Increasingly large genomewide association studies (GWAS) have identified several hundred genomic loci associated with these outcomes, shedding light on additional biological pathways not captured by traditional risk factors (e.g., arterial wall processes in CAD) and identifying novel potential therapeutic targets.

However, similar elucidation of risk factors for less common cardiovascular conditions, such as subarachnoid haemorrhage (SAH), intracerebral hemorrhage or thoracic aortic aneurysm, has not been possible as typical cohort studies have been too small to accrue sufficient numbers of these outcomes. Meta-analyses of such outcomes have typically been small (e.g., ~800 SAH events, Feigin, *Stroke*, 2005) and been unable to implement comparable analyses (e.g., confounder adjustment) across studies due to reliance on previously published results. These outcomes have also remained largely unstudied in GWAS, or have had limited findings due to the small numbers of cases and resultant lack of statistical power (Table 1).

With the establishment of the IHCC and its very large prospective cohort studies, there is an opportunity to rapidly accelerate our understanding of the etiology of less common cardiovascular outcomes, by conducting standardized analyses within cohorts and pooling results across cohorts. Building on our preliminary pilot project on SAH, which so far involves nearly a dozen IHCC cohorts, we request support to scale up and formalize our efforts within IHCC to evaluate the genetic and non- genetic risk factors for 3 further uncommon cardiovascular outcomes over the next two years. The scientific impact will be a dramatically enhanced understanding of the causes of these conditions across a diverse global landscape, informing efforts to treat and prevent disease, including identification of new potential therapeutic targets for pharmaceutical companies. Within IHCC, the impact will include production of a number of ‘quick-win’ manuscripts enabled by the IHCC framework, as well as the development of a collaborative framework for subsequent analytical projects involving existing data from multiple large prospective cohorts.

## 2. Abstract

Background and rationale

The Cardiovascular Epidemiology Unit in Cambridge has a long history of pooling data from prospective cohort studies to provide robust evidence on the association of risk factors with cardiovascular diseases outcomes. Over the past 15 years, through international consortia such as the Emerging Risk Factors Collaboration, we have thoroughly assessed in large-scale pooled data the associations of major lipids (*JAMA* 2009; *JAMA* 2012), anthropometry (*Lancet* 2016; *Lancet* 2014; *Int J Epi* 2012; *Lancet* 2011), diabetes/glycemia (*JAMA* 2015; *JAMA* 2014; *NEJM* 2011), blood biomarkers (*Lancet Diab Endocrin* 2016; *Lancet* 2010 [x3]; *JAMA* 2009), and alcohol (*Lancet*, 2018; *BMJ* 2018) with common cardiovascular outcomes (i.e., coronary disease and ischemic stroke). We have also led (or co-led) international genetics consortia that have discovered several dozen genetic loci associated with these outcomes (e.g., *Nat Genet* 2018; *Nat Genet* 2017).

Idea

We propose to use our expertise in the epidemiology of cardiovascular outcomes to extend beyond major outcomes to uncommon outcomes (e.g., SAH, thoracic aortic aneurysm, pulmonary arterial hypertension), where the risk factor profile has not been well characterized and etiology is not so well understood. Proof-of-

principle projects in this area include our recent evaluation of cardiovascular risk factors for venous thromboembolism in 76 prospective cohorts (*JAMA Cardiology* 2019) and our ongoing project examining genetic and non-genetic risk factors for SAH. For this nascent project we have formed a collaboration, so far including results from ten cohorts (EPIC, Million Veteran Program, UK Biobank, deCODE, Kaiser Permanente-Gera, HUNT, BioVu, BioME, Michigan Genomics Initiative, Geisinger MyCode) involving ~4000 SAH cases.

Impact

Preliminary data from our pilot project on hemorrhagic stroke subtypes for both classical vascular risk factors (**Table 2**) and genetic variation (**Figures 1 and 2**) suggest there is potential for great impact from the proposed projects. Firstly, by combining data from just two of the IHCC cohorts (EPIC and UK Biobank) we have identified that some risk factors (sex and diabetes) are associated with risk of SAH and intracerebral hemorrhage in opposing directions, while other risk factors (e.g., age, smoking) have substantial quantitative differences (**Table 1**). This has implications for prevention and clinical management of these distinct subphenotypes, as well as for the future analysis of epidemiological studies, which have traditionally grouped these etiologically distinct outcomes. Our discovery of the first robustly associated genetic locus for SAH on chromosome 6 (as well as additional highly promising loci that we anticipate being confirmed with additional data contributions from other IHCC cohorts) paves the way for identification of novel etiological pathways, suggesting new potential therapeutic targets. Powerful GWAS results will also facilitate "two-sample" Mendelian randomization studies of classical and novel risk factors, for which well-powered trial data rarely exist.

For each of the uncommon cardiovascular outcomes we propose to study, the IHCC project would be the largest to examine the genetic and non-genetic risk factors and hence there is similarly great opportunity to have a substantial impact on our understanding of disease. Although relatively uncommon, these outcomes still convey a substantial global disease burden (e.g., 500,000 people die annually from SAH, around half of whom are below the age of 50, leading to a similar loss of disability- adjusted life years for SAH as ischemic stroke in the US) so many patients could ultimately benefit.

Timeline

For SAH, where phenotype definitions and statistical analysis plans have already been agreed and results from some cohorts have already been shared, the project can be completed and two manuscripts (one on classical cardiovascular risk factors and one on genetic risk factors SAH) can be produced within 6 months from initiation. For the other 3 outcomes, the anticipated timeline is:

0-3 months: identify and iterate ICD codes for phenotype definition; agree statistical analysis plans

3-9 months: invite cohorts to collaborate, disseminate statistical analysis plans and prosecute cohort-specific analyses

9-15 months: clean results and meta-analyse across cohorts

16-24 months: interpret findings and draft manuscripts for submission

Estimated funding required

~$350,000 including:

- funding for 2 junior post-doctoral researchers in the central analysis team (one epidemiologist who will focus on phenotype definitions and time-to-event analyses of non-genetic risk factors; one genetic statistician who will focus on GWAS and downstream follow-up analyses, e.g., pathway analyses)
- funds to cover data management and analytical costs within cohorts (estimated at $2000 x 25 cohorts)
- data access charges
- publication fees
- travel/conference fees

## 3. Additional Information

a) Why does this project require multiple large cohorts?

For genetic analyses, it is possible to create case-control studies of uncommon outcomes since genetic risk factors are not affected by reverse causation and are less susceptible to confounding than non- genetic factors. However, ascertaining large numbers of patients with uncommon conditions to create powerful case-control studies is laborious and inefficient. For study of non-genetic risk factors, incident cases from longitudinal studies are required in order to avoid reverse causation, recall biases or confounding by treatment that may be associated with presence of disease at recruitment. For these reasons, we believe IHCC is an excellent platform for study of genetic and non-genetic risk factors for uncommon conditions as:

- Very large cohorts are required to accrue sufficient numbers of cases for reliable analysis. For example, we identified >300 incident cases of SAH within the 500,000-person UK Biobank study. By contrast, within the 140 smaller prospective cohort studies in our 2 million-person Emerging Risk Factors Collaboration, no single study had accrued more than 100 incident cases, preventing reliable within-cohort analyses;
- IHCC has wide-spread availability of coded health outcomes data using International Classification of Disease codes within the IHCC cohorts, allowing creating of standardized phenotype definitions across cohorts.
- Information on classical cardiovascular risk factors, such as age, sex, adiposity, blood pressure, blood lipids is commonly available
- Genotype data has already been collected (or will soon be collected) for a substantial number of very large cohorts within IHCC, affording power for genetic discovery analyses
- The IHCC cohorts cover a broad range of ethnicities and continents, which for some outcomes will permit comparison of risk factors across diverse populations, an element of epidemiology that is currently lacking for many conditions.

We propose to invite all cohorts with relevant information recorded (i.e., classical risk factors, ICD- coded outcome information, genetic data) to participate, resulting in sample sizes of several million participants and several thousand events for each outcome.

b) What kinds of data/sample access would be required?

In parallel with the work of Team A to create solutions for data harmonization and the potential for federated analyses, we will proceed with a 'traditional' disseminated analysis strategy. We will therefore not require access to raw individual-level data from each cohort, but transfer of results from analyses to be performed by each cohort according to common analysis plans. This proposal does not require access to samples, but will instead rely on use of pre-existing measurements (e.g., genotype and biomarker data as well as traditional risk factors) made previously by the contributing cohorts.

c) What additional assays or data collection would be required?

For the reasons outlined above, additional assays or data collection will not be required. However, extensions to available genotype data within cohorts from other IHCC or cohort-specific initiatives would enhance the power of the study.

d) What is the data analysis plan?

The broad analytical strategy is for the central analysis team to create phenotype definitions for each outcome (predominantly based on ICD codes or other readily available phenotypic information), write analysis plans for dissemination to contributing cohorts, and set up infrastructure for results sharing (e.g.,

secure FTP sites with cohort-specific logins). Once results are fed back to the central analysis team, data cleaning and harmonisation will be undertaken before results are meta-analysed across cohorts. We are confident that this approach will work well as it has been tested through our pilot analysis of SAH and hence we already have experience of overcoming the challenges (e.g., different study designs).

For the GWAS analysis, typical analysis strategies will be employed (i.e., logistic regression of case/non-case outcomes, adjusting for age, sex and principal components of ancestry, having removed closely related participants). For non-genetic risk factors, we will derive a bespoke set of potentially relevant exposures for each outcome. Risk factors will be tested in a time-to-event analysis using Cox proportional hazards models having excluded participants with cardiovascular diseases at baseline. Sets of potential confounding factors will be pre-defined for consistency across cohorts and will allow participation of the maximum set of cohorts (e.g., one analysis model will include only non-blood-based variables to avoid precluding cohorts without blood measurements). Meta-analyses will typically involve fixed effects for genetic variants (where discovery power is a more significant factor than effect heterogeneity) and random effects for non-genetic risk factors, where quantification of effect estimates and heterogeneity between them is important.

e) What are the technical and/or policy challenges?

The main technical and policy challenge is the current barriers to the sharing of harmonized individual- level data that would permit a centralized analysis in a single IT environment. We propose to work closely with Teams A and C to help overcome these barriers and/or create workarounds, since the proposed project described in this RFI would greatly benefit from central data availability, which would dramatically speed up these multi-cohort analysis projects. Under the existing collaborative framework the main challenge would be incentivising cohorts to conduct the proposed analyses. We are therefore proposing to resource the project to cover data management and analytical costs for contributing cohorts so that local costs are covered, which might be especially important for cohorts coordinating in LMICs.

Tables and Figures

**Table 1. Numbers of cases and loci discovered by previous largest GWAS of proposed uncommon cardiovascular outcomes for study**

| **Outcome** | **Cases in largest GWAS** | **Number of loci found** |
|---|---|---|
| Subarachnoid hemorrhage | N/A | N/A |
| Thoracic aortic aneurysm | ~700 | 1 (*FBN1*) |
| Pulmonary arterial hypertension | ~2000 | 2 (*SOX17*, HLA) |
| Aortic valve stenosis | ~2500 | 3 (*LPA*, *PALMD*, *TEX41*) |
| Chronic thromboembolic pulmonary hypertension | N/A | N/A |

**Table 2. Associations of traditional vascular risk factors with incident subarachnoid hemorrhage and intracerebral hemorrhage**

| | Subarachnoid Hemorrhage | | Intracerebral Hemorrhage | |
|---|---|---|---|---|
| | HR (95%CI) | P-value | HR (95%CI) | P-value |
| Age (per ten years) | | | | |
| EPIC-CVD | 1.15 (0.99 -1.33) | 6.06E-02 | 2.30 (2.05-2.59) | 2.19E-40 |
| UK Biobank | 1.24 (1.07-1.45) | 5.10E-03 | 2.29 (1.92-2.74) | 7.35E-19 |
| Random effects | 1.20 (1.08-1.33) | 1.04E-03 | 2.30 (2.09-2.54) | 1.72E-56 |
| | | | | |
| Sex | | | | |
| EPIC-CVD | 1.74 (1.37-2.23) | 9.91E-06 | 0.70 (0.59-0.83) | 4.95E-05 |
| UK Biobank | 1.86 (1.46-2.38) | 8.97E-07 | 0.78 (0.62-0.97) | 2.53E-02 |
| Random effects | 1.80 (1.52-2.14) | 6.84E-11 | 0.73 (0.63-0.83) | 5.38E-06 |
| | | | | |
| Smoking (current versus never) | | | | |
| EPIC-CVD | 2.66 (2.06-3.42) | 1.79E-13 | 1.43 (1.17-1.73) | 3.76E-04 |
| UK Biobank | 3.22 (2.43-4.27) | 3.42E-15 | 1.42 (1.00-2.03) | 5.08E-02 |
| Random effects | 2.89 (2.40-3.49) | 3.41E-26 | 1.43 (1.20-1.69) | 5.40E-05 |
| | | | | |
| Diabetes | | | | |
| EPIC-CVD | 0.81 (0.40-1.64) | 5.67E-01 | 1.35 (0.95-1.92) | 8.87E-02 |
| UK Biobank | 0.82 (0.45-1.52) | 5.43E-01 | 1.69 (1.18-2.42) | 4.26E-03 |
| Random effects | 0.82 (0.51-1.30) | 3.98E-01 | 1.51 (1.17-1.94) | 1.36E-03 |

**Figure 1. Manhattan plot of interim GWAS of ~2500 cases of subarachnoid hemorrhage**

First locus for SAH identified on chromosome 6, other promising peaks on chromosomes 3 and 9 to be confirmed with inclusion of additional data.

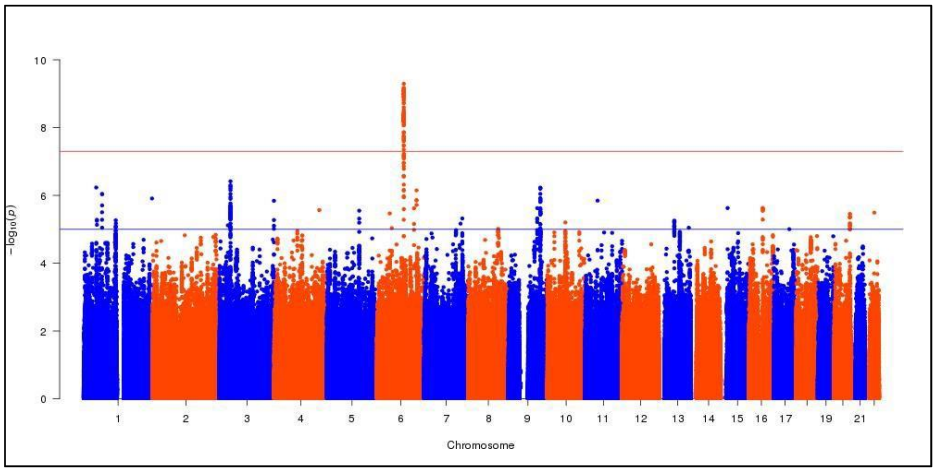
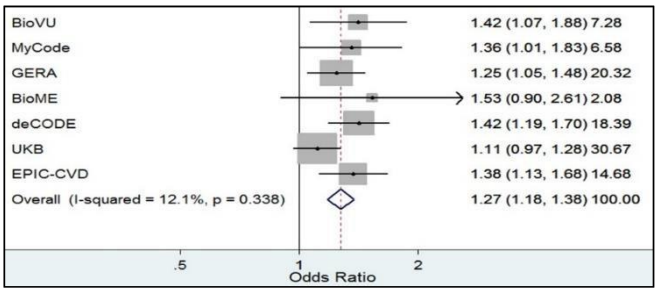



**Figure 2. Forest plot of lead variant associated with subarachnoid hemorrhage on chromosome 6**



# High-throughput Metabolomic Biomarker Measures Across IHCC Cohorts

**Investigators:** Adam Butterworth, Ph.D.; John Connolly, Ph.D.; Hakon Hakonarson, M.D., Ph.D.

**Primary Cohort Sites:** University of Cambridge Cohort (laboratory of Dr. Butterworth); Center for Applied Genomics Cohort at Children's Hospital of Philadelphia (CAG Cohort @CHOP, laboratory of Dr. Hakonarson)

## 1. Idea Summary

To improve risk prediction and prevention of chronic diseases, such as diabetes and heart disease, we propose a low-cost metabolic profiling technology that measures 228 biomarkers for lipids (including 14 lipoprotein subclasses), particle size, apolipoproteins, fatty acids, omega-3 and -6, amino acids, ketones, chronic inflammation, fluid balance, kidney function, and glycolysis. This captures far more the handful of biomarkers by routine cholesterol tests, and at a relatively small proportion of the cost. At a scale of more than 100,000 individuals, it can be available at <$20 USD or <€18 euro.
The platform has several advantages in addition to its expanded biomarker panel – it allows for highly stable repeatability of measurements, in absolute concentrations and with no batch effects. High-throughput metabolic profiling can be used to provide sample quality assessments and can scale to any number of samples. Nightingale's technology has applications in various areas of medicine with particular relevance in epidemiological research.

The Children's Hospital of Philadelphia (CHOP) and University of Cambridge will partner with Nightingale Healthcare (Finland) to improve risk prediction and prevention of chronic diseases in children. Piloting Phase 1 in the U.K. and U.S., our goal is to initially pilot the approach on 10,000 patients, where electronic health record (EHR) and biobank data are already linked.

Phase 2 – beyond scope of the current proposal – would be pioneered with developing countries where access to (hitherto) prohibitively expensive technologies and services, has resulted in a dearth of genomic data derived, and disproportionately incomplete understanding of disease etiology and risk factors.
Our focus is pediatric disease in minority groups. At CHOP, we have multiple collaborations with groups in South-America and China, while our own pediatric cohort is enriched for children with African American ancestry (~39% of ~100,000 patients). The University of Cambridge, meanwhile, has extensive collaborations with groups in SE-Asia.

## 2. Structured Abstract

Background:
Beyond sample collection, the Nightingale's service handles every process step from handling to blood sample analysis to health data delivery. The biomarker analysis service is based on nuclear magnetic resonance (NMR) spectroscopy, and can provide biomarker measurements from any serum or plasma sample that has been collected and kept according to standard procedures for lipid testing (300 ul required, but can go as low as 100 ul at a slightly higher cost). Blood metabolites are quantified in absolute concentrations (e.g. mmol/l) and percentages, so that biomarkers can be analyzed like any other clinical chemistry data. They have been used previously to improve risk prediction and biomarker discovery of cardiovascular diseases, type 2 diabetes and other disease targets, with a multitude of alternate applications available.
Idea: The blood analysis service has the ability to stratify patients, taking into account the individual differences in phenotype, a crucial next step towards achieving a more personalized approach. In addition to supplementing existing approaches to biomarker identification, the low-cost technology offers

the opportunity for previously-excluded areas of relative impoverishment to participate in high-level biobank-driven science with immediate clinical application.

Impact:
Heart disease, diabetes and other chronic diseases often develop without symptoms, resulting in heavy burdens for both patients and healthcare systems. Sufferers of chronic diseases can struggle with disease management and disease progression can lead to severe health complications.
The proposed program enables early identification of disease risk through metabolic profiling, empowering patients to take better control over their own health. By enabling early identification of the onset of chronic diseases we can also effectively target medications, prescribing treatments to the right patient groups and avoiding overmedication.

Further, the ability to clinically track human health at both the patient and population level provides transformative possibilities for governments to build effective nationwide health policies. Rapid and accurate disease risk profiling predicts health outcomes and allows responsive feedback on the effect of drug and lifestyle intervention strategies.

From a pharmaceutical perspective, combining a lipid panel with multiple other metabolite groups to provide novel molecular insights into the effect of lifestyle factors, genetic makeup and drug mechanisms. In the preclinical phase, metabolic profiling helps predict drug effects, while in clinical phases it can be used to establish personalized treatments.

Metabolic profiling enhances target identification and validation, even at the preclinical stage. Combining metabolic profiling with genomics using biobank cohorts can reveal early molecular insights into drug target biology and characterize pleiotropic effects. Nightingale's high-throughput platform can also be used to discover novel drug candidates through PheWAS approaches.

Timeline:
Year 1: Pilot:
- Run on 5,000 pediatric patients at CHOP
- Run on 5,000 patients at University of Cambridge

Year 2:
- EHR Data integration and analyses across both sites (n=10,000)
- Genotype integration with all participants
- Risk factors analysis
- Phase 2 applications submitted based on Preliminary Data

Year 3: Complete PRSs for:
- Publication of data
- Continuation to Phase 2, incorporating sites from Africa, South Asia, and/or South America.

Estimated funding:
$400,000 (Phase 1)

**3. Additional Information**

Why does this idea require multiple large cohorts? Which cohorts would be required?
The program offers a high-yield return on investment in supplementing existing healthcare data and provides an opportunity to bring developing countries into the fold as beneficiaries of individualized medicine.

Required data/sample access:
Genetic data

- Genotypes

Phenotype data

- EHR-derived or standardized measures preferable
- Medication history

Additional assays or data collection required:
We do not anticipate need for additional assays.

What is the data analysis plan?
Analysis of the associations between metabolic measures and relevant phenotype (e.g. lipid levels) using univariate linear regression and between genetic variants and the phenotype following a conventional GWAS approach, will be performed. Two different albeit complementary models will be used:

1. Standard analysis of association between metabolic metabolite levels (such as lipid profiles) and genetic data

A conventional GWAS approach will be used to test for associations between genetic variants and the phenotype of choice. The analysis is performed following the standard single-SNP approach where SNPs are tested one at a time. Associations will be investigated using linear regression assuming an additive effect on the trait and including sex as a covariate. We will use linear regression to test SNP effects on the serum metabolites, such as the measured lipid profiles in the context of multiple disease phenotypes, such as cardiovascular disease and diabetes. Analyses will be performed following a univariate approach where metabolites are tested one at a time. We will include sex as a covariate to correct for sex differences in serum metabolic profiles. To identify significant associations between individual SNPs and serum metabolites in patients with different phenotypes, we will adopt a conservative Bonferroni-corrected significance level, $p < 0.01/M$, where M denotes the total number of serum metabolite measures.
Analyses will be conducted including all the individuals for whom genetic, metabolic and phenotype data are available (N = 10,000).

2. Genome metabolome integrated network analysis

This method consists of two stages: (i) construction of the differential network, and (ii) a genome-wide correlation analysis (GWCA). We start by performing a differential network analysis that allows us to test whether the pattern of pairwise associations between metabolites is the same in subjects with different disease, such as cardiovascular disease and diabetes and whether it significantly differs across groups. To eliminate the confounding effect of sex on the serum metabolites, the data used for this analysis are the residuals from a linear regression model of each metabolite on sex.
Briefly, the underlying interdependencies between metabolites are initially measured for each of the physiological groups using shrinkage estimates of partial correlations. To test whether the association between metabolites significantly differs between groups, we perform a two-sample permutation test. We will use 100 000 permutations in our analysis. If the partial correlations between two given metabolites are significantly different between the two physiological groups, then we draw an edge in the differential network. The connections included in the differential network are defined by setting a cut-off on the two-tailed p-value. The power to estimate correlations is lower than the one to estimate a change in mean levels, therefore, to infer the differential network we will set an uncorrected threshold, $p < 0.01$. To validate the differential network analysis results, we compare the network structure between the cohorts. The replicated results between cohorts are further investigated in the next step of the analysis.
In the second step, we perform a GWCA to identify genetic variants associated with differences in metabolic associations. As for the standard GWAS study, all individuals for whom genetic data are available will be included in the analysis. To find the desired associations, we first classify individuals according to the number of copies of the less frequent allele carried, giving genotype groups A (0 copies) and B (one or two copies). SNPs are tested one at a time. Subsequently, the correlation between metabolites m1 and m2 is calculated for each group, rA and rB, and differences in correlation between groups A and B is also tested. As in the differential network analysis, we eliminate the confounding effect

of sex on the serum metabolites by using the residuals of a linear regression of the metabolite level on sex. To test whether the two correlation coefficients, rA and rB are the same, we use the z transform method as described in reference. To correct for multiple testing, significant associations between genetic variants and variations in metabolic associations are determined using the genome-wide significance threshold $p < 5 \times 10-8/D$, which corresponds to a genome-wide significance level adjusted with the number of differential connections (D) identified in step one. To assess the biological significance of our findings, identified SNPs are assigned to the nearest gene (maximum distance 1 Mb).

Technical and/or policy challenges that will need to be addressed:
Data-sharing across cohorts.

# Identification of Loss of Function (LOF) Variants in Health and Disease

**Lead Investigator:** Dong Li, Ph.D.

**Primary Cohort Site:** Center for Applied Genomics (CAG) Cohort of the Children's Hospital of Philadelphia (CHOP, laboratory of Dr. Hakon Hakonarson)

## 1. Idea Summary

Identification of loss of function (LoF) variations as risk factors for chronic disease represents an important area of research in genomic medicine, providing a rubric for (individualized) screening and risk prediction, as well as tools for exploring functional/etiological disease mechanisms. Pairing the study of LoF variations with a discrete phenotype (e.g. lab value) represents a tidy experimental design, but requires large sample-set for requisite power. For example, several LoF variants in PCSK9 have previously been associated with reduced levels of LDL cholesterol and heart disease, prompting several high-profile translational studies of PCSK9 inhibitors. We have identified several cases of PCSK9 knockout in obese children with no effect on triglyceride levels. Building upon these findings, we aim to further examine the impact of PCSK9 LoF and deletions in diverse and minor populations across several large cohorts, which have direct implications for drug development, prescribing, and population screening. In addition, we will also focus our analysis on other genes known to harbor null mutations as well as candidate genes where null mutations would have robust impact similar to PCSK9, including but not limited to APOC3 in LDL-C, ANGPTL4 in coronary artery disease, ANGPTL3 in LDL-C, and HSD17B13 in chronic liver disease. Importantly, focusing on rare LoF variants and deletions association analysis in large cohort will give us an opportunity to discover novel genes across multiple diseases and related intermediate traits, with the ultimate goal of fully understand the phenotypic consequences of LoF in every gene in the human genome. Due to the genetic heterogeneity for most (if not all) of diseases and human traits, we also aim to leverage consanguineous populations, in which the gene pool derives from a small number of population founders owing to bottleneck effect and genetic drift, to identify homozygotes for enriched LoF and deletion alleles.

## 2. Structured Abstract

Background and rationale:
Understanding the biological function of every gene in the humans represents a primary goal of the biomedical field. While rare genetic variants predicted to disrupt biological function (so-called LoF) have been traditionally studied in the context of severe Mendelian diseases, Mendelian randomization studies of genetically determined complex human traits (e.g. LDL cholesterol and triglyceride levels) have suggested that rare trait-associated LoF variants are actually causal. For example, gain-of-function mutations in proprotein convertase subtilisin-kexin type 9 serine protease (PCSK9) have long been associated with high LDL-C, atherosclerotic cardiovascular disease (ARCVD), and familial hypercholesterolemia (FH), while loss of function mutations are associated with decreased risk (1, 2). The translational implications of these studies have led to several clinical trials with PCSK9 monoclonal antibodies (3), and in 2015 the Food and Drug Administration of the United States approved the PCSK9 inhibitors alirocumab and evolocumab for treating FH and ARCVD (though access to PCSK9 mAbs remains problematic).

Idea:
Our group at CAG/CHOP has identified several pediatric obese patients with PCSK9 LoF variants have high triglyceride levels. We aim to further examine the impact of PCSK9 LoFs (including deletions in PCSK9) in extreme distributions across several large and diverse cohorts, also focusing on identifying disease modifier genes/variants through sequencing of the subjects with LoF variants. We will also pursue other genes with null alleles that are protective to human diseases. Additionally, putting together all the samples allow us to identify novel genes (knockouts) that are associated with or causal for human

diseases and related traits, especially by leveraging the founder populations, where rare founder alleles (in other out-bred populations) can increase in frequency.

Impact:
This will broaden our understanding of the biological function of human genes, which have direct implications for drug development, prescribing, and population screening.
Timeline:
3 years in total. We would anticipate 6-12 months to obtain the access to the diverse worldwide cohorts. CNV and variants calling on genotyping and sequencing data would take an additional 1-3 months. Association analysis would take an additional 6-12 months. Then we would anticipate functional studies on candidates to take an additional 8-12 months.

Estimated funding required:
$100,000-150,000 per year. Pending funding status, cost-sharing could be considered.

**3. Additional Information**

Why does this idea require multiple large cohorts? Which cohorts would be required?
Systematic analyses of large genome datasets will provide sufficient statistical power to overcome experimental and biological noise. While the project can go forward with fewer cohorts, in addition to our CHOP cohort, we would like to require access to the following cohorts:

23andme, US
45 and Up study
Africa Centre for Health and Population Studies, South Africa
BBMRI-NL-Biobank
Biobank Japan
BioVU Vanderbilt
China Kadoorie Biobank
China PEACE
Chinese Newborn Sequencing Project
Danish National Biobank
Danish National Birth Cohort
East London Genes and Health
Estonian Genome Project
Genomics England / 100,000 Genomes Project
Kaiser Permanente Research Program
Korea Biobank Project
Korean Genome and Epidemiology Study
German National Cohort
Golestan Cohort Study
LifeGene (and sister cohort, EpiHealth), Sweden
Million Veteran Program, US
MyCode Community Health Initiative, US
Norwegian Mother and Child Cohort Study
UK Biobank
UK Blood Donor Cohorts
UK Collaborative Trial of Ovarian Cancer Screening
PERSIAN Cohort Study
PROMIS
Saudi Human Genome Program
Saudi National Biobank

Required data/sample access:
Genetic data (sequence, genotypes, or lipid panel) and BMI. Phenotypes such as LDL scores, cardio and liver phenotypes would also be useful.

Additional assays or data collection required:
CHOP and UPenn have extensive resources and models available for use in functionally evaluating and validating candidate causal variants, including, but no limited to:

1. Genetic mouse core facility for creating knockouts, knock-ins, and transgenics
2. zebrafish core facility that can readily create knock-downs of specific genes and generation of specific gene mutations
3. Drosophila facility with library of genetic mutants and assays for assessing phenotype
4. Core facility for generation of induced pluripotent stem cells (iPSCs) from patient-specific samples and ability to manipulate iPSCs genetically using targeting strategies using TALENs and CRISPRs, and AAV vectors.
5. Creation of lymphoblastoid cell lines from patient-specific peripheral blood samples.
6. Expression of recombinant forms of genes containing candidate causal variants in various types of standard cell lines (e.g. 293T, CHO, lymphoblastoid cell lines).
7. Criteria for which candidate causal variants to pursue functional validation
    1) causal variants that are potentially actionable, i.e. amenable to therapeutic intervention
    2) availability of relevant in vitro and/or in vivo models for functional validation
    3) novelty of the implicated gene, impact on phenotype, and/or mechanism of action

What is the data analysis plan?
CAG is a world leader in creating genome analysis software, having developed some of the most widely-used tools in academia, such as ANNOVAR (4768 citations) and PennCNV (1342 citations). For structural analysis, we have unparalleled expertise, with >50 relevant publications (of >600 total). We propose to perform CNV calling on all the accessible genotyping cohorts and LoF analysis across the datasets including genotyping and sequencing data as standard. We have developed a comprehensive pipeline for the annotation and eventually interpretation of genomic data. Our pipeline identifies known and novel disease-causing variants in a high throughput fashion, prioritizes them for follow-up, and relates them to existing clinical data.

Analysis of sequencing data: the first step employs a parallelized read mapping and variant-calling pipeline that leverages existing tools; e.g., Burrow Wheeler Alignment (BWA). This pipeline is API and Amazon cloud compatible, and hence is capable of calling SNVs and indels in a rapid, highly parallelized fashion. Also included are several innovative methods for optimization and quality control of the variant calling procedures used in this step of the pipeline. NGS technologies are producing unprecedented quantities of data to be processed into biologically relevant information. The computational and statistical tools required for the task have been lagging behind the technology development resulting in a continual evolution of mapping and variant calling tools. We routinely benchmark new tools against our established pipelines to compare the mapping and variant calling accuracy and speed of new tools, updating the pipeline as warranted. The key components of the variant-calling pipeline are: 1) Pre-processing and quality control of raw reads outputted from sequencing instruments 2) Mapping: align the read against the Human Genome Project reference 3) variant calling and annotation: identification of SNPs, indels and CNVs across the dataset using a multi-sample caller and d) local realignment: re-calibration and re-alignment of reads at indels.

Analysis and interpretation of copy number variants (CNV) (microarrays and sequencing):
Copy number analysis (with emphasis on homozygous deletions) will be performed with similar strategies for chromosomal microarrays and WGS data. Each alteration containing a CNV involving one or more genes or 9 or more SNPS will be analyzed for potential clinical significance. The genetic content and frequency of each CNV will be reviewed using information from several databases including OMIM, Database of Genomic Variants, ClinVar, DECIPHER, Gene Reviews and PubMed and correlated to the

patient's phenotype. We will compare the data to an internal database of over 450,000 thousand patient chromosomal microarray data.
Analysis of LoF variants: Nonsense, frameshift, canonical splice-altering mutations, or CNV are predicted to inactivate a gene. To increase the probability that mutation correctly annotated as predicted LoF (pLoF) by automated algorithms, we will remove nonsense and frameshift mutations occurring within the last 5% of the transcript and within exons flanked by non-canonical splice sites, splice-site mutations at small (<15 bp) introns, at non-canonical splice sites, and where the purported pLoF allele is observed across primates. Common pLoF alleles are less likely to exert strong functional effects as they are less constrained by purifying selection; thus, we define pLoF mutations as variants with a minor allele frequency (MAF) of <1% and passing the aforementioned bioinformatic filters.

Test for association of LoFs and CNVs: While we will initially focus on associations of LoFs alleles with the various human diseases and related traits, our study design will allow for careful dissection of structural variations (including deletions) in the germline of our study subjects. The realization that the normal human genome contains regions that can vary in copy number from individual to individual was an unexpected and exciting finding. Multiple investigators have suggested that this form of variation might be responsible for some of the variation in disease expressivity that is seen in patients with genetic disease. We hypothesize that copy number variation in risk alleles is a susceptibility factor for various traits we will be examining.

Technical and/or policy challenges that will need to be addressed:
Data-sharing across cohorts.

**Bibliography**


1. Abifadel M, Varret M, Rabes JP, Allard D, Ouguerram K, Devillers M, Cruaud C, Benjannet S, Wickham L, Erlich D, Derre A, Villeger L, Farnier M, Beucler I, Bruckert E, Chambaz J, Chanu B, Lecerf JM, Luc G, Moulin P, Weissenbach J, Prat A, Krempf M, Junien C, Seidah NG, Boileau C. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. Nat Genet. 2003;34(2):154-6. Epub 2003/05/06. doi: 10.1038/ng1161. PubMed PMID: 12730697.

2. Cohen JC, Boerwinkle E, Mosley TH, Jr., Hobbs HH. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N Engl J Med. 2006;354(12):1264-72. Epub 2006/03/24. doi: 10.1056/NEJMoa054013. PubMed PMID: 16554528.

3. Baum SJ, Cannon CP. PCSK9 inhibitor valuation: A science-based review of the two recent models. Clin Cardiol. 2018;41(4):544-50. Epub 2018/03/08. doi: 10.1002/clc.22924. PubMed PMID: 29512936; PMCID: PMC5947644.

# International 100k+ Consortium (IHCC) Drug Development Resource

Aroon Hingorani, FRCP, Ph.D., Director, UCL Institute of Cardiovascular Science, UCL Professor of Genetic Epidemiology and Consultant Physician, UCL Hospitals NHS Foundation Trust (For the UK Longitudinal Women's Cohort)
Usha Menon, M.D., M.B.B.S., Professor of Gynaecological Cancer, MRC Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology, Faculty of Population Health Sciences (For the UK Longitudinal Women's Cohort)
J.P. Casas, M.D., Ph.D., Scientific Director of Partnerships, Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Healthcare Boston, Executive Director, Million Veteran Program, U.S. Department of Veterans Affairs (VA), US

## 1. Idea Summary

The international 100k+ consortium is in a prime position to support drug development and repurposing, whereby information on drug targets and disease is derived from human genomic and linked biomedical data analysed at scale. By helping to de-risk drug development through human genomics, the consortium has the opportunity to reduce the high rate of late stage failure. Ensuring industry works on the correct therapeutic targets for each disease should help stimulate work on new mechanisms of action and diseases that have previously been considered too risky to tackle, ensuring development pipelines become more focused on therapeutic need, to the benefit of patients and societies globally.

## 2. Structured Abstract

### a. Background and rationale

A critical task in drug development is to identify the key proteins ('therapeutic targets') that are involved in disease pathogenesis and which are also amenable to therapeutic action, usually by small molecule drugs or monoclonal antibodies. Traditionally, this task has been informed by pre-clinical laboratory experiments in (cell, tissue, and animal) disease models but these are notoriously poor at predicting clinical efficacy in the intended indication, contributing to the high rate of late-stage drug development failure.

### b. Idea

Genetic association studies in patients and populations test relationships between natural sequence variation (genotype) – assayed at low cost by genotyping arrays, or more expensively by sequencing – and disease biomarkers or clinical end-points (phenotype) – many ascertained through research, but thousands of others through linkage to routinely recorded healthcare data. Genetic effects (like drug action) are mediated through proteins (according to Crick's Central Dogma), and variation in the genome is inherited at random (according to Mendel's Laws), much like treatment allocation in a clinical trial. Thus, variants in a gene encoding a drug target, that alter its expression or function, can be used as a tool to anticipate the effect of drug action on the same target. Numerous proof of concept examples now provide evidence for the utility of this paradigm, some incorporating complementary information from other 'omics (e.g. proteomics and metabolomics), The approach can be used to deduce biomarker profiles that signify drug target engagement, distinguish on- from off-target actions of drugs, identify indication expansion or repurposing opportunities, anticipate mechanism-based adverse effects, and even success or failure in late phase clinical trials. The paradigm also makes it feasible, for the first time, to match drug targets to disease end-points systematically and comprehensively (target identification), through genome wide association studies (GWAS). GWAS test variants in all genes against a single phenotype, with more stringent control over the false positive rate than is usual in most preclinical laboratory studies ($p \leq 5 \times 10\text{-}8$ vs $p \leq 5 \times 10\text{-}2$). Phenome wide association analysis (PheWAS – in which variants in a gene encoding a druggable target are tested against not one but many diseases and biomarkers) complement GWAS by helping to anticipate the effects of drug action beyond the primary disease indication (target validation). Adding proteomic, metabolomic, physiological and imaging data to

genomics can help uncover potential mediating pathways. Therefore, use of genomics to match targets to diseases, predict the diverse beneficial and adverse mechanism-based effects of drug action on any target, and illuminate mechanism, addresses the major cause of clinical phase drug development failure. Many cohorts in the IHCC have connected (or have the potential to connect) genotype to disease biomarkers and clinical end-points at scale, to provide a rich resource for drug target identification and validation for human disease. This proposal is to develop genomic data-driven drug target identification and validation as a key IHCC initiative, with standardized approach to data generation, meta-analysis, reporting, visualization and sharing, in a form that is useful for drug development.

c. Impact

The currently high failure rate in drug development is reflected in the high price of new drugs and, in turn, in cost pressures on healthcare systems. This places a financial burden on citizens (through health insurance premiums or taxation) and, increasingly, limits access to new drugs because they exceed cost-effectiveness thresholds set by payers. Many diseases remain poorly treated because they are considered too risky to tackle. Many drugs which were shown to be ineffective in their intended indication, but which have proven safe in man, have been shelved, but could be repurposed if the correct indication(s) for their target could be found. The genomic (data) driven approach to drug development directly addresses all these issues. Reducing high rates of clinical failure due to lack of efficacy should contain drug development costs and increase accessibility to future medicines. Finding new uses for safe but failed drugs would increase the range of available treatments. Greater drug development efficiency would stimulate work on diseases previously considered too risky to tackle. Since genetic association studies often implicate the same gene (and therefore drug target) in several diseases, there is also the potential to design treatments for several conditions in tandem, to help address the growing problem of multi-morbidity.

d. Timeline

Since many of the tasks needed for this initiative are the same, or similar to, those established for large-scale genetic association studies for disease gene discovery, which many IHCC cohorts have already participated in, useful knowledge for drug development could emerge within a 2-3 year time frame. A key task will be to harmonise phenotypes and disease end-points across studies, and to enrich key studies with genotypes where samples are available but DNA extraction and genotyping has not yet been undertaken. Through its links with funders, and the potential to make pre-competitive links with multiple industrial partners, IHCC is in a strong position to create a dataset of several million participants from multiple cohorts internationally to execute this work.

e. Estimated funding required

This depends on the scale of the initiative and whether new genotyping and new ‘omics measures are in scope. More detailed costing would be required if the initiative is prioritized by the consortium.

**3. Additional information**

a. Why does this idea require multiple large cohorts? Which cohorts would be required? [suggest that the idea submitter reach out to the cohort leaders to discuss the idea and seek their input and approval]

Cohorts are ideally designed to achieve the aims of this proposal because (unlike case-collections) they accrue information on multiple disease end-points, and frequently make measures of pre-clinical disease biomarkers, including new proteomics and metabolomics measures. However, the success of the proposed initiative rests on linking genotype to phenotype with scale, breadth and depth that has not been achieved before. Scale because small genetic effect sizes and stringent control over false discoveries necessitate very large sample sizes. Although a few meta-analyses of genetic association studies have included $>10^6$ participants, most have been insufficiently powered to discover all the available druggable targets. The discovery of even small genetic effects is valuable because drugs can be developed to target the same mechanism with effects that are 6-10 times larger than that of the

corresponding genetic variants. Breadth because GWAS to date have investigated only a fraction of the known human diseases. Of the 10,000 or so diseases found in the International Classification of Diseases -10th revision, only a few hundred have been studied by GWAS. Resolution because advances in proteomics, metabolomics and imaging are making it possible to better sub-stratify patients and diseases and understand mechanism but, with a handful of notable exceptions few of the available datasets have incorporated such measures at scale.

b. What kinds of data/sample access will be required?

The initiative would require access to genotype, disease biomarker and clinical end-point data.

c. What additional assays or data collection will be required?

For some cohorts with stored biological material new DNA extraction and genotyping would be required. New proteomics and metabolomics measures on stored samples would also be needed in some cohorts to help inform mechanism of action and to help reliably assign genetic associations to specific targets.

d. What is the data analysis plan?

The analyses proposed could be achieved through a federated meta-analysis approach without any transfer of participant level date. Basic code and scripts in R are available for executing such analyses, and could be adapted for the initiative.

e. What are the technical and/or policy challenges that you anticipate will need to be addressed for this idea to be successful?

An information governance and data sharing model would need to be agreed but a federated meta-analysis approach has proven successful in analyses of this type in the past and obviates the need for any transfer of participant level data. The way in which the consortium interacts with industry needs careful consideration. Models would need to be developed that recognize the contribution of citizens' genomic and healthcare data to the value chain, such that this is reflected in the pricing and accessibility of new drugs developed through knowledge contributed by the consortium.

# New Atlas of Genetic Influences on Human Blood Metabolites Based on Single Nucleotide Polymorphisms and Structural Variants

**Author:** Christian Gieger, Ph.D., on behalf of the German National Cohort (NAKO) OMICS group

## 1. Idea Summary

We intend an improved calling of structural genetic variants by long-read sequencing in large world-wide cohorts from Europe, Africa, America, and Asia. These new, unique data will be used to complete the atlas of genetic influences on human blood metabolites.

## 2. Structured Abstract


a. *Background and rationale*
Nowadays sequencing of samples in population-based cohorts is standard and large collections are on their way. The technique has shown its advantage over genotyping in finding not only new rare but also common single nucleotide polymorphisms (SNPs) and short insertion or deletion polymorphisms (indels). But due to using these short-read sequencing techniques important structural genetic variants have been ignored so far.

b. *Idea*
Long-read sequencing allows detecting large numbers of new unique structural variants of all types and new indel variants. Moreover these new data lead to an improved phasing of structural variants together with single nucleotide variants into allele-specific haplotypes. In this project we use the German National Study (NAKO) together with studies from Europe, Africa, America, and Asia to sequence high quality DNA of hundred of samples from each study. Based on these sequences we will create new reference genomes for these worldwide populations that reveal several mega-bases of new sequence parts and include several thousands of new structural variants (Ameur A et al, Genes (Basel). 2018 Oct 9;9(10)).

c. *Impact*
These new, unique data will be used to complete the atlas of genetic influences on human blood metabolites (Ref: A genome-wide perspective of genetic variation in human metabolism, Illig T, Gieger C, et al. Nat Genet. 2010 Feb;42(2):137-41; Human metabolic individuality in biomedical and pharmaceutical research, Suhre K, et al. Nature. 2011 Aug 31;477(7362):54-60; An atlas of genetic influences on human blood metabolites, Shin SY, et al, Nat Genet. 2014 Jun;46(6):543-550). This work will be followed by a multi-omics approach to develop new biomarkers for cardiometabolic traits and diseases. In parallel the project will provide the obtained data for other genetic epidemiological project. The next obvious step is to validate the structural variants and to check the relevance of the derived markers in a clinical setting by using standard sequencing approaches.

d. *Timeline*
The project will use the improvement in speed of generating long-read sequencing in the next two years. Our aim is to generate 100 long-read sequences for each of six cohorts with different ethnicities. In total we expect to need about one year to generate the intended 600 sequences.

e. *Estimated funding required*

This project will extensively use the drop of prices for long-read sequences in the next two years. We expect a price of 1.500€ per sample in near future. This sums up to in total 900.000€ for 600 samples. In addition we estimate to need 100.000€ for the measurement of approximately 1000 metabolites in 600 plasma samples using 2 complementary metabolic platforms. In total the project needs a budget of 1.000.000€.

## 3. Additional Information

a. *Why does this idea require multiple large cohorts? Which cohorts would be required? [suggest that the idea submitter reach out to the cohort leaders to discuss the idea and seek their input and approval]*
Human genetic variations are the genetic differences in and among populations. A variety of studies have indicated that GRCh38 may be incomplete. This incompleteness has different characteristics for specific population groups. Of all new sequences found by long-read sequencing in a Swedish population only about 60% of these novel sequences are shared with a Chinese genome (Ameur A, et al , Genes (Basel). 2018 Oct 9;9(10). With this project we start to depict the genetic diversity between different ethnicities based on newly found SNPs and structural variants. We plan to sequence 2-3 European cohorts and 2-3 Asian cohorts, if available also one African and one American cohort.

b. *What kinds of data/sample access will be required?*
We require participation of population-based cohorts representing all populations of their regions. We select a random sample of 100 participants of each of these selected cohorts. The cohorts should be able to extract high quality DNA and they should have blood plasma available.

c. *What additional assays or data collection will be required?*
We plan to sequence DNA by using a long-read sequencing platform (e.g. PacBio) and measure metabolite patterns with 2 metabolomics platforms (e.g. Metabolon and Nightingale Health)

d. *What is the data analysis plan?*
Quality control and genome alignment will be based on proprietary software of the chosen sequencing platform and open source software (e.g MUMmer3). Summary statistics for the filtered whole genome alignments will be calculated with R packages. Long read structural variations will be detected with a combination of Sniffles and NGMLR. BLAST comparisons will help to annotate new sequences. For analyses of associations between genetic variations and metabolite levels standard linear regressions will be used.

e. *What are the technical and/or policy challenges that you anticipate will need to be addressed for this idea to be successful?*
We expect no insuperable technical or policy challenges

**Members of the German National Cohort (NAKO) OMICS group are**

Christian Gieger, Helmholtz Center Munich, Germany; Thomas Illig, Hannover Medical School; Harald Grallert, Helmholtz Center Munich, Germany; Anna Köttgen, University of Freiburg, Germany; Iris Heid, Regensburg University, Germany; Kathrin Günther, Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany; Georg Homuth, University of Greifswald, Germany; Wolfgang Lieb, University of Kiel, Germany; Tobias Pischon, Max Delbrück Center for Molecular Medicine, Berlin, Germany; Dan Rujescu, University of Halle, Germany; Tanja Zeller, University of Hamburg, Germany;

Börge Schmidt, IMIBE, Essen, Germany, Hans Grabe, University of Greifswald, Germany; Anna Flögel, Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany; Tobias Kerrinnes, Helmholtz Centre for Infection Research, Braunschweig, Germany; Eva Reischl, Helmholtz Center Munich, Germany; Norman Klopp, Helmholtz Center Munich, Germany; Andre Franke, University of Kiel, Germany; Matthias Nauck, University of Greifswald, Germany; Christian Müller, University of Hamburg, Germany; Hermann Brenner, University of Heidelberg, Germany; Eleftheria Zeggini, Helmholtz Center Munich, Germany; Astrid Petersmann, University of Greifswald, Germany; Markus Löffler, University of Leipzig, Germany; Markus Noethen University Hospital Bonn, Germany; Markus Scholz, University of Leipzig, Germany; Annette Peters, Helmholtz Center

# Non-adherence to Medical Tests and Treatments and Future Mortality

**Author:** Paul Pinsky, Ph.D.

## 1. Idea Summary

This project attempts to define a health behavioral profile of non-adherence to medical tests and procedures and examine the relationship of this profile with future mortality.

In the context of cancer screening trials, non-adherence to protocol screening has been shown in three trials to be associated with a large increase (50% or more) in mortality over the subsequent decade from causes unrelated to the screening. The idea for the current project is to see whether such an association can be investigated in the context of routine clinical practice, outside of a research trial setting.

With cancer screening, it is easy to separate the direct effects of the test being non-adhered to, which would only affect deaths from the cancer being screened for, and indirect effects, which are responsible for the observed increase in other-cause (not related to the screened-for cancer) mortality. The indirect effects could be two-fold. First, non-adherence could be correlated with other health behaviors, such as smoking or diet (say, resulting in obesity) and these factors could affect future mortality. Second, the non-adherence with cancer screening could be associated with a general non-adherent health behavioral profile, in that those with the profile are more likely to forego tests and treatments in the future, whether of a preventive nature or not, and that this behavior leads to increased mortality. In the cancer screening trial, even after controlling for smoking and obesity (and other factors), there was still an almost 50% increase in other-cause mortality over the next decade.

The idea here is to develop a health behavioral non-adherence profile, based on observed adherence within a given age range to recommended medical work-ups, tests and procedures related to primary care and prevention. With this non-adherence profile, which defines adherent and non-adherent (and partially adherent) cohorts, one would examine its association with future mortality. Further, given an observed increase in mortality associated with the non-adherence profile, one would try to delineate the causes as to direct effects of the tests that were missed, effects related to correlations with other health behaviors (e.g., smoking), and effects of a general non-adherent behavioral trait.

## 2. Structured Abstract

Background and Rationale

A recent paper showed that non-adherence to cancer screening exams was associated with a substantial increase over the following 10-15 years in mortality unrelated to the cancers being screened for (Pierre-Victor et al., JAMA Int Med, 2018). This finding was among participants in a U.S. randomized trial of cancer screening (the PLCO trial, one of the cohorts in the 100K+ Cohorts Consortium). Because they volunteered for the trial and consented to receiving screening, they were presumably healthy enough to receive the tests. The magnitude of the increased risk for overall mortality (excluding from trial cancers) was large, 73% over ten years (controlling for age, sex and race), and remained large, 46%, even after also controlling for behavioral factors and comorbidities. A similar finding was observed in cancer screening trials in Italy and the U.K. The excess risk in PLCO was of similar magnitude to that of morbid obesity. About 10% of subjects were non-adherent even in this population with a healthy volunteer effect. In the U.K. and Italian trials, 30-40% were non-adherent.

These findings were in the context of subjects enrolled in randomized trials. It is of interest to see if similar findings could be observed in the context of routine clinical care. Other studies have examined adherence to treatments for specific diseases and morbidity and mortality for those diseases. However, to our knowledge, there is little or no literature on the relationship between adherence to standard preventive medical tests and visits in the context of routine care and future mortality.

b. Idea

The idea is to define a non-adherence profile based on medical tests and treatments and health-care provider encounters. This would be based primarily on routine checkups, well visits and recommended preventive tests. A target age for the assessment of a non-adherence profile would be around 45 to 54 years. In this age interval, individuals generally are healthy and regular checkups and preventive services are recommended. However, it is a young enough range for interventions to begin that could substantially impact future morbidity and mortality.

The primary questions are as follows:

1) Can one develop a non-adherence profile based on physician visits and medical tests received during, say, the decade of 45-54 in generally healthy individuals based on available EHRs in a number of the cohorts?
2) If so, is this non-adherence profile associated with increased all-cause mortality over the succeeding decade or two?
3) Can any increase be separated into the effect of missing the non-adherence profile tests per se, the confounding effect of correlated health behaviors (e.g., smoking), and the effect of a general non-adherence profile (manifesting itself with future non-adherence to medical tests and treatments).

If the answers to the first two questions are yes, then the next step would be to try to develop interventions to identify and target these individuals and attempt to alter their non-adherence behavioral profile, through health education, counseling, etc.

c. Impact

Given the large magnitude of the observed effect of non-adherence to medical tests on overall mortality, and its relatively high frequency, the potential public health impact of a successful intervention is substantial. For example, given a 20% prevalence and an RR of 1.5, the population attributable risk is 9%. For overall mortality, that is very large. For comparison, the population attributable risk for smoking in the U.S. has been estimated at 20%.

d. Timeline

The first step would be working with EHR data across cohorts and deriving common data elements; this could take 4-6 months. The next step would be the actual analysis, which could take another 4-6 months. Note this assumes there are cohorts that already have follow-up of 10-15 years, as well as the requisite data to define non-adherence.

e. Estimated funding required

Funding would be required for programmers and data analysts to work with the EHR data across different cohorts, and statisticians and epidemiologists to perform the analysis. There is no funding required to collect or analyze specimens or to collect additional data.

**3. Additional Information**

a. Why does this idea require multiple large cohorts? Which cohorts would be required?

This idea requires multiple large cohorts for several reasons. First, it is of interest to know whether this finding is robust over various settings and populations, for example, across different racial/ethnic groups, across different types of health care systems, etc. Second, to guard against overfitting, it would be useful to have separate validation cohorts, where the non-adherence profile is fixed beforehand. Third, large numbers of deaths are required to examine different broad categories of causes of death (e.g., cancer,

digestive diseases, respiratory diseases). Finally, the non-adherence profile definitions may need to be different by sex, and possibly other factors (race/ethnicity, rural/urban status), requiring larger numbers of individuals.

Based on the cohort profiles, the following cohorts could have the necessary data for the project based on linkage to EMRS and long enough follow-up for mortality: 45 and Up Study ; AMORIS; California Teachers Study; East London Genes and Health; Estonian Genome Project; EPIC; Generations Study; Genomics England; Kaiser Permanente Research Program on Genes; Nurse Health Study; Ontario Health Study.

b. What kinds of data/sample access will be required?

No access to samples is needed. For data, access to individual patient level data would be required.

c. What additional assays or data collection will be required?

No addition assays are needed. No additional data collection should be required.

d. What is the data analysis plan?

For the data analysis, the first step is to attempt to define a profile for non-adherence to medical tests and procedures. This would involve the following: a.) choosing which medical tests, treatments and health care provider encounters go into making up the non-adherence profile, and also which age interval to use for when these events occur; b.) based on the defined set of events above, developing an algorithm to define a non-adherence profile variable, with levels of non-adherent, adherent and possibly an intermediate level (partially adherent). These two steps would be stratified by gender.

Both of the above steps should be taken by researchers who are blinded to the outcome status of the cohort subjects (i.e., mortality events after the age range on which adherence is defined).

The second step would be to evaluate the association of the non-adherence variable with future mortality. Follow would begin after the age range during which the non-adherence profile is defined. The analysis would assess excess mortality risk over different time periods, and for various categories of causes of death. Other covariates would also be examined for stratified analyses (e.g., racial/ethnic subgroup, pre-existing conditions, etc.).

The evaluation of the association of non-adherence and mortality could be undertaken in two steps with a training and test (validation) set. Some cohorts or subsets of all cohorts could be held back and the association evaluated in a training set of outcomes. Based on the results, the definition of non-adherence could be tweaked, and then fixed to be analyzed in the remaining validation set of outcomes.

The third step would be to disentangle the direct effects of missing the tests comprising the non-adherence profile per se and the more general effect of a non-adherence profile (manifesting itself in non-adherence to tests and procedures in the future). Note confounding with other health behaviors (e.g., smoking) should be relatively easy to adjust for, assuming the necessary variables are in the cohort data bases. One approach would be isolate sub-components of the non-adherence profile and examine causes of death unrelated to those components.

e. What are the technical and/or policy challenges that you anticipate will need to be addressed for this idea to be successful?

A technical challenge would be to create common data elements based on EMRs/EHRs from different countries and different systems.

# Pharmacogenomic Analysis Across IHCC Diversity Cohorts for Assessment of Drug Response

**Lead Investigators:** Kenny Nguyen, Ph.D., and John Connolly, Ph.D.

**Primary Cohort Site:** Center for Applied Genomics Cohort at Children's Hospital of Philadelphia (CAG Cohort @CHOP, laboratory of Dr. Hakon Hakonarson)

## 1. IDEA SUMMARY

The pharmacogenomics (PGx) between genetic variant and drug response associations have been well-documented from expanding on the single drug-gene (pharmacogenetic) interactions to a more comprehensive approach dealing with the effects of multiple genes from a drug response. Big data and data analytics methods will be explored to elucidate questions to translational research and precision medicine. Here, four major issues are addressed that presently plague PGx progression:

1) For well-known drug-gene pairs, clinical recommendations are absent or inconclusive for understudied or unidentified (rare) alleles;
2) A lack of ancestry-specific data for certain drug-gene pairs;
3) Predictions differ for individuals who are genotyped on different arrays and sequencing platforms for clinical PGx implementation; and
4) A general lack of sophistication in applying novel approaches to study drug effects across populations.
5) It is believed that targeting genomic and electronic heath record (EHR) data from several large cohorts may elucidate new discoveries for a novel informatic pipeline for clinical, healthcare, and medical purposes. Therefore, this proposal satisfies the IHCC Scientific Strategy and Cohort Enhancements (Team B) and its charge to develop a scientific agenda by identifying novel approaches to diagnose and treat genomic conditions, explore enhancements to existing cohorts, and address diversity gaps. In addition, project ideas and key goals are addressed as described from the SSCE to improve understanding of variability in response to treatments, which include the following targets addressing the above issues:

- Identify novel PGx associations
- Cross-cohort genetic array analysis of drug responses
- Identify and understand the response mechanism and metabolism of commonly used drugs

## 2. STRUCTURED ABSTRACT

(a) BACKGROUND AND RATIONALE

Background: The growth rate of biological data has been significantly underestimated [1] and will become problematic as this value becomes greater than the rate available to analyze it (P versus NP) [2]. With proper informatics infrastructure in place, real-time analyses and decisions that are vital for PGx and for proper implementation in a clinical environment are able to move forward.

Rationale: The use of alternative and enhanced high performance and throughput computing methods are hypothesized to further understand the data vacuum being accrued over an extraordinary rate. Big data predictive analytics is one of the two novel focal points addressed here where they will be tested for internal benchmarking and validation. It is important to note that predictive analytics – unlike its descriptive and prescriptive analogs – is a data reductive method that uses various computational, mathematical, and statistical modeling and (deep) machine learning techniques in artificial intelligence to analyze past data and predict future outcomes. In tandem with the DL approach – while conserving the theme of reduction – coarse- graining (CG, or renormalization) will also be attempted to elucidate genomic and EHR data as an attempt to analyze the largest data sets in a tractable manner. CG is a method that has been successful in statistical physics, and has been used to reduce the degrees of

freedom to elucidate the molecular behavior. As a result, calculations on a spatiotemporal scale can be reduced by a factor of three to four allowing simulations to increase by three orders of magnitude.

(b) IDEA
The idea was derived from the need to address vital issues in big data analytics associated to medical genetics and genomic medicine. As mentioned earlier, novel methods must be used to analyze increasing genomic data being stored at various institutions all over the world. Not addressing this issue will create an intractability problem. In terms of analyses, the use of a systems approach to PGx, formulate a CG expression through Gaussian functions, and creating training sets for DL will be done in tandem to benchmark and validate each other, and has recently been implemented towards efforts in drug design, development, and discovery, as well as to further enhanced turnaround time in repurposing, repositioning, and rescue (DRPx).

(c) IMPACT
The results of this proposal will illustrate the impact on enhanced methods in analyzing PGx data with speed velocity and efficiently while maintaining the original integrity of the data due from a reductive approach. Once they are internally validated, they can be tested and benchmarked by other institutions. A systems approach to PGx has already been developed as an extension to biology and pharmacology. The results from the CG and DML approaches will be compared to it.

(d) TIMELINE
The proposed work will take less than or equal to three years.

(e) FUNDING
Funding for the proposed project will be projected at $750,000 over 3 years. Pending availability of funds, cost- sharing could be considered for the successful implementation of this project.

## 3. ADDITIONAL INFORMATION

(a) COHORTS
With relatively smaller sets of data, brute-force calculations can be done with reasonable computational power. However, the methods discussed here would examine very large data sets to execute and observe the success of this concept. Specific cohorts from the IHCC Summary Table have been selected to demonstrate the power and turnaround time of the discussed methods:

- 23andMe
- China Kadoorie Biobank
- Estonian Genome Project
- Genomics England / 100,000 Genomes Project
- Kaiser Permanente Research Program
- Korea Biobank Project
- Korean Genome and Epidemiology Study
- LifeGene (and EpiHealth)
- Million Veteran Program
- MyCode Community Health Initiative
- Norwegian Mother and Child Cohort Study
- UK Biobank
- UK Blood Donor Cohorts
- UK Collaborative Trial of Ovarian Cancer Screening

(b) DATA/SAMPLE ACCESS
Data or sample access of genomic data will be analyzed with an initial focus on genotypes,

sequencing, or targeted panels. EHR-derived data will initially focus on prescriptions, dosing, allergy, and/or adverse drug responses from the institutions mentioned above will be used to create training sets.

(c) ADDITIONAL ASSAYS/DATA COLLECTION
Assays and collection will eventually extrapolate to genomic and EHR data with PGx associations. If necessary, data from additional institutes will be analyzed from the IHCC Summary Table.

(d) DATA ANALYSIS
HPC/HTC: The computational infrastructure for data science (discover/access/distill) and engineering (extract/load/transform) is just as (if not more) important than the data itself. The reality is that storage, pipeline development and optimization for processing, and raw data are dependent to each other. Three different pipelines will be used on PGx-associated genomic and EHR data.

DL: The Dartmouth Conferences in 1956 birthed the field of AI. In the decades since, advances in computing has allowed it and its ML descendant to be the face of AI until 2012. Since 2015, AI has exploded in the form of DL from the big data movement. Much of that precipitated from the wide availability of massively paralleled computing – i.e., graphical processing units – that make processing ever faster, cheaper, and more powerful to render genomic data, images, text, transactions, mapping data, etc. This stems from extracting usable information from large data sets requires computational approaches to pattern-detection prediction, detection, and classification. The co-application of statistical methods and DL provides a powerful and agnostic means of deriving these outcomes and have been proposed to accelerate novel discovery [3].

CG: Genomic and EHR PGx-related data will be harvested and curated for renormalization analysis. This method has already been done for principle component analysis of biological and financial data [4]. CG alleviates the idea to use brute-force calculations in the event of difficulty to observe simulation convergence, which is typically the endpoint for such an event. Convergence can be defined as reaching the energetic minimum. In order to achieve this, important steps are necessary to create a CG model. This is done by decreasing the degrees of freedom (or resolution) in a system. Once this is done, rescaling or renormalization may be necessary in order to increase the smoothness of the fitting Gaussian function. Determining the optimal DOF will take the most amount of time to det
ermine the correlation length through renormalized magnetization to obtain free energy values.
Systems PGx: This is a direct result from extending systems biology and pharmacology. Dynamic mathematical models have already been developed and tested using a variety of biological data sets and will be also applied to PGx data using existing methods [5].

Benchmark/Validation: All de novo model prediction methods will go through a normalized confidence analysis. A threshold will be created to determine scores will be high or low. If they are low, then quality analysis and control will be performed to determine why it did not pass threshold. Conversely, high confidence scores will be promoted to decide on keeping the method.

Case (1): An example of interrogating the first major issue examines the ability of different genotyping and sequencing platforms to detect alleles in 11 well-known “PGx genes” [6]. Importantly, with exome-sequencing (ES), alleles defined by variants in introns or promoters (e.g., CYP2C19*17 and CYP3A5*3), are not interrogated with major obvious implications for pharmacogenetic recommendations for these two alone relevant to 13 medications with CPIC guidelines. To remedy this, it has been proposed to further catalog such discrepancies, as well as genotype with ES-only data.

Case (2): Another example that addressed the second major issue, which involves ancestry and uncatalogued variants. The frequencies of many PGx alleles differ greatly by ancestry, association, and

drug response.
CYP3A5*3 allele has been found at a frequency of 98% in an Iranian population, but at only 11% in a population from Malawi [7,8]. Further exploration by cataloging these differences will be critical to promoting future development and implementation of PGx projects worldwide.

(e) TECHNICAL/POLICY CHALLENGES
Data-sharing across cohorts is always a challenge and will be facilitated with proper IT infrastructure. We will generate a hub for advanced and enhanced methods for elucidating PGx-related data.

REFERENCES


(1) Stephens ZD, Lee SY, Faghri F, et al. Big Data: Astronomical or Genomical? PLoS Biol. 2015;13(7):e1002195. doi:10.1371/journal.pbio.1002195.

(2) Piotr Indyk. 2017. Beyond P vs. NP: Quadratic-Time Hardness for Big Data Problems. In Proceedings of the 29th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA '17). ACM, New York, NY, USA, 1-1. doi:https://doi.org/10.1145/3087556.3087603.

(3) Kalinin AA, et al. Deep learning in pharmacogenomics: from gene regulation to patient stratification. Pharmacogenomics. 2018;19:629–650. doi: 10.2217/pgs-2018-0008.

(4) Bradde S, Bialek W. PCA meets RG. J Stat Phys. 2017;167(3-4):462-475.

(5) Ribba B, Grimm HP, Agoram B, et al. Methodologies for Quantitative Systems Pharmacology (QSP) Models: Design and Estimation. CPT Pharmacometrics Syst Pharmacol. 2017;6(8):496-498.

(6) Reisberg S, Krebs K, Lepamets M, et al. Translating genotype data of 44,000 biobank participants into clinical pharmacogenetic recommendations: Challenges and solutions. Genet Med. 2018;20:1. doi:10.1038/s41436-018-0337-5.

(7) Bains RK, Kovacevic M, Plaster CA, et al. Molecular diversity and population structure at the Cytochrome P450 3A5 gene in Africa. BMC Genet. 2013;14:34. Epub 2013/05/07. doi: 10.1186/1471-2156-14-34.

(8) Rahsaz M, Azarpira N, Nikeghbalian S, et al. Association between tacrolimus concentration and genetic polymorphisms of CYP3A5 and ABCB1 during the early stage after liver transplant in an Iranian population. Exp Clin Transplant. 2012;10(1):24-9.

# Public Health and Genetic Dimensions of Iron: A Proposal to the International 100K Consortium

**Authors:** Robert B. Wallace, M.D., M.Sc., University of Iowa College of Public Health; David Melzer, M.D., Ph.D., University of Exeter

## 1. Idea Summary

Iron is involved in most human and animal metabolic processes, and is said to be related to more human illnesses than any other substance. This proposal would use large cohorts of the IHCC to further explore the epidemiological and biological roles of iron exposure in many important diseases and conditions. These would include studies related to iron deficiency (e.g., anemias and conditions of blood loss, and various comorbid conditions), and to iron overload conditions, such as those potentially related to neurodegenerative diseases (e.g. dementias and multiple sclerosis), heart and liver diseases, the possible increased risk of acquiring various infectious diseases and the demonstration of increased iron burden leading aging syndromes such as sarcopenia, frailty and chronic pain. The exploitation of cohorts with different demographic characteristics, varying risks of diseases and prevalence rates for genetic variants, and a broad range of relevant environmental exposures, including those that are iatrogenic (e.g., thalassemia), could promote a productive global research program leading to a range of disease preventives and population health improvements.

Important knowledge gaps could be addressed in this project, with targeted research studies: What populations and their characteristics are at greater risk of iron deficiency? What dietary iron levels are associated with the organ dysfunction and diseases of iron overload? Can some of the degenerative diseases of aging be mitigated by decreasing iron exposure or treating excess levels? Do elevated iron levels "travel" with other environmental exposures that can be controlled? Are there genes related to iron absorption, metabolism and homeostasis that can help explain some iron related conditions in specific individuals or populations? Can practical iron-related screening programs be conducted in both clinical and population settings? These are questions that can be addressed by diverse, large, well-documented cohorts.

## 2. Structured Abstract

*a. Background and Rationale:*

While an enormous amount of useful work has been conducted on iron and its molecular congeners over the past few centuries, there are still enormous gaps in knowledge about the role of iron in normal physiology and how its regulation and homeostasis, including genetic and epigenetic factors, relate to various kinds of important clinical conditions, including iron-deficiency anemia. The genetic origins of hemochromatosis have importantly increased our understanding of iron-related genetic processes and the role iron overload on organ function and diseases. Yet, this condition is still a public health problem; screening is uncommon and many cases are not diagnosed until organ disease or failure is uncovered. The work of Melzer and colleagues has recently extended the role of iron overload to important geriatric syndromes, suggesting possible new control measures. The genetics of iron metabolism now also extends to the regulation of iron homeostasis, the structure and function of many iron-containing proteins, and the redox functions of iron accumulation that may have an important pathogenic role in many additional conditions. Indeed, the genetics of iron molecular and metabolic processes extends beyond humans to infectious agents, such as malarial parasites and invasive bacterial species that adversely affect the human host.

*b. Idea:*

The basic "idea" is contained in the Idea Summary above. The following are examples of <u>selected *specific aims*</u> that would be of great interest to us, depending on the nature (data and determinations already in place) and accessibility of the cohorts, as well as the interests of cohort investigators:

1. Determine the role of dietary iron (including supplements) and on geriatric syndromes such as frailty, sarcopenia and falls, and on related infectious disease occurrence.
2. Determine the role of iron intake on targeted organ diseases known to be associated with increased iron intake and accumulation, such as in the heart, brain, liver and skeletal muscle.
3. Assess the feasibility of screening for body iron burden (e.g., serum iron, ferritin, etc) and genetic markers of increased iron burden.
4. Determine the prevalence of the iron deficiency anemia in various demographic and the role of pre-specified genetic variants related to iron absorption and homeostasis.

*c. Impact*:

All good research depends on the ultimate findings. We believe that important public health and clinical findings can emerge from this program in the prevention and control of iron-related health problems.

*d. Timeline:*

This program is envisioned to be three years in duration. Later extension could occur depending on study findings, and cohort resources and data availability.

*e. Funding Requirements*:

At this moment, budget planning is difficult. We expect that this project will require about $250,000 (US) per year to conduct the breadth of studies envisioned, depending on cohort data access and content. Funds would cover a small portion of senior investigator salaries, including Wallace, Melzer and statistical/ data management staff. Other funds would devoted to data analysis, communication with cohort staffs, cohort data acquisition, laboratory determinations where feasible, and acquiring informed consent where necessary.

**3. Additional Information**

The approach to this proposed research program, if approved and funded, would be to gather investigators affiliated with relevant cohorts, and identify and plan for a set of specific aims that that can be performed in the planned 3-year time frame (see above) and are compatible with cohort characteristics, including clear study documentation, necessary sample sizes, specimen and biomarker availability, risk factor measurement, clinical data availability, particularly on outcomes responding to the scientific hypotheses. The cohorts in this consortium (including some in which we participate “locally,”) are large and complex, and acquiring documentation, data dictionaries and datasets will take some time.

*a. Why large, multiple cohorts?*

Our hypotheses could include cohorts with different characteristics, such as presence of young adults as well as elders, geographic and racial/ethnic diversity, cohorts with documented and reasonably accurate clinical outcomes and biomarkers, various genetic and physiological determinants already in place as well as available stored genetic and blood materials (and possibly pathology specimens as available) to conduct ancillary determinations. *Which cohorts?* As noted in our cover letter, we only learned of this consortium a few days before this submission, and cohort identification and contact is not feasible, but would take place immediately after submission of this proposal.

b. *Kinds of data/sample access required*.

In general, we would need general data documentation and data dictionaries for the chosen cohorts. We would also need also need relevant, requested clinical and research laboratory determinants, and the availability of stored samples, mostly blood and/or DNA, but possibly tissue pathology samples as well. For certain studies we would need certain clinical outcomes from clinical records. We appreciate the sensitivity and data acquisition processes that might be required here, but for most of the studies, self-reported outcomes would not be suitable.

c. *Additional assays required*.
In addition to those mentioned above, we would need genetic determinations for genes of interest. There may be circumstances where whole genome sequencing would be of great value, as would some epigenetic determinations; the latter have been associated with regulation of iron status. We understand that these determinations might not be available, but under some circumstances we may be able to perform genetic and epigenetic determinants within limits of resources.

d. *Data Analysis Plan*.
Data analysis will be performed in two venues. The analysis of risk factors and health outcomes within cohorts, such as by using proportional hazards and other survivorship methods, as well as summary statistics of categorical and continuous variables, will be conducted by senior staff biostatisticians from the Department of Biostatistics at the University of Iowa College of Public Health. Genetic analyses will be performed by Dr. Melzer and his genetics team at the University of Exeter. Since we are not presenting specific studies at this moment, detailed analytical plans are not feasible, but there is ample experience in both groups to conduct all modern analytical methods.

e. Technical and Policy Challenges
There are a number of potential challenges to this project. We anticipate that some portions of various cohort data will have inadequate documentation. Some cohorts where data collection is still in progress may not be able, understandably, to deliver data or specimens in a timely manner. There are potential problems with data interoperability that will likely be resolved, but this often takes time. In a few instances, particularly when new determinations are proposed, there may be challenges in obtaining informed consent. Similarly, some countries may have objections to the transfer of tissue/blood specimens, genetic information or other physiologic determinations. This will need to be anticipated and in some instances certain proposed studies may not be performed because of these impediments. Shipping logistics can sometimes be an important problem. Finally, while there is great utility in cross-translating and harmonizing both questionnaire data and laboratory determinations, this can require considerable time and resources, and it is hoped that the Consortium will facilitate these activities.

# Rare Recurrent Copy Number Variants (CNVs) in Health and Disease

**Investigator:** Joseph Glessner, Ph.D.

**Primary Cohort Site**: Center for Applied Genomics Cohort at Children's Hospital of Philadelphia (CAG Cohort @CHOP, laboratory of Dr. Hakon Hakonarson)

## 1. Idea Summary

A comprehensive map of rare single nucleotide and copy number variants is a critical gap to progress in disease association studies. Observing rare homozygous recessive and lack of expected homozygous recessive according to Hardy-Weinberg equilibrium for those not tolerated for life is critical in assessing the common disease rare variant model as well as the missing heritability gap left by common allele genome-wide association studies. Quality control of variant classifications is key, including evaluation of no-call genotypes which may be mis-clustered rare homozygous variants. Imputation to bolster the number of variants and normalize the data format and scaling is critical to make the many datasets comparable. Meta-analysis of large sample sizes with an evolving disease model of different ancestral background, and with better refinement of genes and sub-regions, including leveraging summary statistics when raw genotyping is not available, yields a foundation for optimally powered analysis. Almost 200 million genotyped samples provide an unprecedented look at the genetics of the world's populations with respect to ethnicity, sex, age, and disease status. Our CNV platform, ParseCNV2, was designed to convert array PennCNV and sequencing VCF into plink bed format for efficient association and querying. Mega-analysis combining multiple orthogonal lines of evidence to implicate biologically plausible disease mechanisms in the context of are and ultra-rare variants, may give unprecedented insights into disease biology. Tracking of potentially confounding factors such as batch effects as covariates will lead to a properly conditioned analysis to infer disease risk and relevant prevention measures across multiple cohort sites.

## 2. Structured Abstract

**a. Background and rationale**
Genome-wide association studies of common single nucleotide polymorphisms have yielded many risk factors for disease and enhanced our biological understanding of manifestation of multiple disease states. However, rare variants have been more difficult to place into proper context due to limited sample sizes and related technical issues of not being able to cluster or predict the expected mode of rare variants which had not been observed in the small cohort on the order of thousands of cases and thousands of controls.

**b. Idea**
IHCC provides an unprecedented opportunity to allow cross talk between the world's genotype data to understand the rare and ultra-rare variant spectrum overall and in various ethnic groups. Quality control and data normalization will be key factors which we have addressed with customized and adaptive software tools which can operate at scale efficiently. Informatics to allow for various formats of genomic and genetic data promotes data inclusion and comparability. Rare alleles or lack of rare alleles where they would be expected due to Hardy-Weinberg equilibrium could be comprehensively cataloged in a database hosted online for internal access and external controlled access.

**c. Impact**
Rare heterozygous and homozygous variant observations in key disease genes can bolster the list of reportable variants to patients, allowing for an end to diagnostic odysseys for patients with both rare and ultra-rare disorders. These discoveries may also help unravel the underlying molecular mechanisms of more common disorders. Being able to filter putative variants from an exome sequencing experiment will be much easier with a comprehensive database allowed by IHCC analysis.

**d. Timeline**
We estimate a two-year timeline to gather, normalize, quality control, and analyze the data.

**e. Estimated funding required**
We estimate $100,000 per year of funding to be required for computational staff and resources ($200,000 total). Pending availability of funds, there may be opportunities to cost-share.

## 3. Additional information

a. Why does this idea require multiple large cohorts?
Which cohorts would be required? [suggest that the idea submitter reach out to the cohort leaders to discuss the idea and seek their input and approval]
Rare and ultra-rare variants require multiple large cohorts to be observed or inferred based on imputation or rare variant clustering methods. As many cohorts as would be willing to participate at whatever capacity they can would be accepted.

b. What kinds of data/sample access will be required?
The flexibility of the informatics approach allows for different kinds of genetic and genomic data to be leveraged. Two-dimensional (genotype allele frequency and intensity) data underlying the genotype as well as the called genotype would be ideal but summary statistics would also work if there are restrictions for data sharing. We anticipate only a few sample aliquots would be needed to verify by an independent technology the presence of a few highly interesting rare variants as a proof of principle that the methods are working well.

c. What additional assays or data collection will be required?
We do not think additional assays or data collection will be required to achieve the aims of our idea to capture the rare variant spectrum in worldwide populations. We will leverage the existing data.

d. What is the data analysis plan?
Gather, normalize, quality control, and analyze the data using Aspera file transfer and ParseCNV2 optimal tools for the work. Fisher's exact test and rvtests software will capture the significance of various rare variant frequencies. No call genotypes will then be investigated to assess their probability of being a false negative rare allele genotype call.

e. What are the technical and/or policy challenges that you anticipate will need to be addressed for this idea to be successful?
The intensity data needed to infer copy number variants may not be available from all contributing sites. The data may be on different scales and in different formats that need to be normalized, which we have anticipated in our software design to accept partial or different data and allow it to contribute in an unbiased way to the insight gleaned. The authorship policy among so many data sources could be a challenge, taking into account size of the cohort but also contribution to significant findings. As all data are de-identified we do not anticipate any unexpected policy challenges.

# The Human Knockout Project

**Author:** Sekar Kathiresan, M.D.; Amit Khera, M.D.; Eric Minikel; Daniel MacArthur, Ph.D.; and David van Heel, Ph.D.

## 1. Idea Summary

Despite many decades of impressive progress in human biology, three-quarters of human genes have no known associated phenotype, and at least a third of them are of completely unknown function. Naturally occurring human genetic variation presents a powerful lever into this uncharacterized biology. Traditional *forward genetics* - aggregating individuals with unusual phenotypes, and characterizing the underlying genetic basis - has already revealed nearly 4,000 genes linked with monogenic disease, and thousands of genomic regions associated with complex disorders such as type 2 diabetes. With the development of very large biobanks with genome-scale DNA sequence and clinical phenotype data, we now have an unprecedented opportunity to leverage natural variation for *reverse human genetics* - identifying individuals with “extreme genotypes” and investigating their biology.

We propose the aggregation of DNA sequence data from hundreds of thousands of humans from diverse biobanks and cohorts to develop a **Human Knockout Project**: an international effort to identify rare individuals with inactivating mutations in genes of interest, ultimately cataloguing whole-organism *in vivo* models of heterozygous and homozygous inactivation for the majority of protein-coding genes. Methodically recontacting these individuals for extensive phenotyping will yield a database of loss-of-function variants compatible with life, and an entry point into the physiological changes caused by disruption of each gene. This database would increase our understanding of the function of human genes, and convert uncharacterized genes into biological insights and validated drug targets. It would also provide a pilot for larger-scale genotype-guided recontact experiments across international cohorts.

This project leaders have extensive experience in large-scale identification and interpretation of loss-of-function variation (Daniel MacArthur), and recruitment and study of consanguineous individuals (David van Heel; Sekar Kathiresan); as well as junior investigators with expertise in variant interpretation and clinical phenotyping (Amit Khera) and the application of genetic variants to drug target validation (Eric Minikel).

## 1. Structured Abstract

Reverse genetics—the engineering of specific genetic changes, followed by observation of the resulting phenotypes—has played a key role in investigating gene function in model organisms. While genetic engineering is not an appropriate approach in humans, human reverse genetics is possible using naturally occurring genetic variation. Given known mutation rates, we can infer that any specific possible single-base substitution compatible with embryonic survival almost certainly exists somewhere among the seven billion living humans. This also means that **for all genes where disruption is compatible with survival to birth, there exist hundreds to thousands of individuals carrying heterozygous—and, in many cases, homozygous—variants that result in complete loss of protein function**. Identifying and deeply phenotyping these individuals would allow us to **characterize the functions of the large majority of human genes *in vivo*.**

We propose, as an initial phase of the Human Knockout Project, to use only existing exome and genome data from IHCC cohorts with appropriately consented participants. We will then use the existing data to identify homozygous loss-of-function mutations of interest, recontact and phenotype those individuals and build a database to share the resulting information with the scientific community.

This project will involve:

- **Producing a catalogue of high-confidence loss-of-function variants** from exome and genome data from the cohorts selected above, using both automated annotation tools and manual curation of variants
- **Developing the ethical, financial, logistical and technical framework required for large-scale global recontact experiments**, including guidelines for appropriate reconsent, procedures for broad phenotyping of recontacted individuals, and centralized collection of phenotype data. We will review the consent, data availability, and demographic parameters across IHCC cohorts to identify the cohorts who are best suited for participation in this project.
- **Recontacting sequenced individuals worldwide who are heterozygous or homozygous for loss-of function variants** in genes for which loss-of-function phenotypes are currently unknown, with a special focus on genes falling within GWAS signals for phenotypes of medical interest.
- **Consenting and then phenotyping those individuals** via questionnaires and standard medical procedures, for features related to the gene in question.
- **Developing a public database of loss-of-function variants**, including aggregate-level associated phenotypes and availability of biobanked specimen and health record data

This 3-year pilot for the Human Knockout Project will greatly increase our understanding of human gene function in an unprecedented way, will likely yield novel drug targets for a variety of indications, and will create a framework for recontact and re-phenotyping of individuals globally that could be used for many other genotype-based questions. We estimate that due to the infrastructure development required, and the labor-intensive process of recontacting study participants that this project will cost $2.2M/year for those 3 years. We foresee **a second, larger stage of the Human Knockout Project that would involve the selection and sequencing of autozygous individuals from currently unsequenced IHCC cohorts**, thereby creating the largest possible database of human knockouts and their associated phenotypes.

**Additional information**

*a. Why does this idea require multiple large cohorts? Which cohorts would be required?*

True loss-of-function variants are typically extremely rare, and often largely or entirely restricted to specific populations, meaning that the power for discovery of true “knockouts” for most genes is very low even in large outbred populations. Two strategies exist for increasing power to discover complete genetic “knockouts” of a specific gene.

The first strategy is to recruit cases from historically bottlenecked populations, in which a loss-of-function variant in the target gene has become common through a founder effect; in such cases, any given gene has a small probability that one or more loss-of-function variants have become common in any specific population, so assaying knockouts in a large sample of human genes will require collections of samples across many populations.

A second, and typically more efficient, strategy for identifying human knockouts is to focus on populations with high degrees of consanguinity (parental relatedness), which contain many individuals with high degrees of autozygosity. Autozygous individuals have a much higher yield of homozygous rare variants than outbred populations. However, cross-cohort analyses will nonetheless be required to increase the overall number of genes with at least one knockout identified.

We will initially focus on cohorts with existing sequencing data (or with plans for sequence data generation over the next three years); with considerable existing clinical and phenotype data collected; with appropriate consent for genotype-based recontact; and, ideally, with demographic properties that increase the probability of an increased yield of homozygous loss-of-function variants. Model cohorts for this purpose include the East London Genes & Health (ELGH) cohort, led by PI van Heel, which samples a population (individuals of Pakistani origin living in the U.K.) enriched for consanguinity, linked to electronic medical

record and other trait data, and with full consent for genotype-based recontact; and the PROMIS cohort, based in Pakistan, another recontactable cohort with high consanguinity for which PI Kathiresan has already led the generation of over 40,000 exomes or genomes. Other large cohorts we have previously worked with include the Mount Sinai (BioMe) Biobank, the Estonian Biobank, and multiple Finnish cohorts aggregated through the FinnGen consortium. Our ultimate goal in this proposal will be to **investigate the full range of IHCC cohorts** to expand this study to all suitable populations.

*b. What kinds of data/sample access will be required?*

We will require access to raw sequencing data from participating cohorts, which will then be reprocessed and jointly-called to produce a single harmonized collection of variants (as we have previously done with gnomAD). Following the discovery of candidate LoF variants, we will request access to de-identified clinical and phenotype data from individuals identified as homozygous or compound heterozygous for such variants. Data types that we have experience with handling include electronic medical records, health questionnaire data, and biomarker measurements.

Overall, we will request and harmonize such data for at least 1,000 participants per year, including LoF homozygotes and matched controls, across all of the participating biobanks.

*c. What additional assays or data collection will be required?*

For a small number of individuals (at least 50/year) with homozygous variants in genes of special interest, we will request recontact for deeper phenotyping, as well as matched controls from the same cohorts. We will prepare a standardized multi-system health questionnaire to be administered to all individuals investigated in this project, to be translated into additional languages as needed. We will request basic laboratory tests, including serum assays of lipids or other biomarkers. Where possible, we will request serum and cells or tissue from the individuals, which will enable validation of the functional impact of the candidate loss-of-function variant, as well as downstream analysis of biological impact. In addition, we will perform focused phenotyping experiments, as described below.

*d. What is the data analysis plan?*

Raw sequencing data will be reprocessed and jointly called to create a unified variant call file. Variants across all of the cohorts will be annotated using the LOFTEE tool, which identifies candidate gene-disrupting variants and removes multiple common classes of annotation artifact, and improves the detection of candidate splice-disrupting variants. Candidate LoF variants that pass LOFTEE filters will then be manually assessed by trained curators using a custom visualization portal that allows rapid inspection of raw read data from variant carriers, as well as deep annotation data including gene models from RefSeq and Ensembl, evolutionary conservation, and gene and transcript expression across tissues; we have already used a pilot version of this portal to curate more than 4,000 candidate loss-of-function variants.

We will identify high-confidence loss-of-function variants in genes of strong biological interest, particularly those in genes of unknown biological function that have been implicated in human disease or traits through unbiased genetic analyses (e.g. genome-wide association studies), or genes for which inhibitory therapeutics are currently being actively considered. We will also prioritize genes for which multiple independent loss-of-function variants are available, ideally spanning multiple cohorts, to improve the value of this project as a testing ground for cross-cohort analysis. Once homozygous carriers of those variants have been identified, we will request available already-existing clinical and other phenotype data from the cohorts, both for the homozygous carriers and (where possible) for a large set of controls matched for age, sex, genetic ancestry and cohort. We expect to request such existing phenotype data for at least 1,000 individuals per year.

For approximately five genes of highest biological interest each year, we will initiate full recontact experiments, requesting additional phenotyping of the relevant homozygous carriers; we expect to recontact at least 50 individuals per year. Proposed phenotyping will combine unbiased health scans (multi-system health questionnaires, health record harmonization and analysis, comprehensive metabolomics) with focused phenotyping developed in collaboration with gene-specific domain experts (imaging for subclinical abnormalities, perturbation studies that unmask an underlying phenotype). An important recent example is prospective phenotyping of individuals with protective homozygous loss-of-function variants in the dietary fat clearance pathway (in the *ANGPTL3* gene), where we noted absence of the protein product in circulating blood, absence of coronary plaque on imaging studies, and a remarkable ability to clear fat following a dietary challenge (Stitziel *et al.* 2017, J Am Coll Cardiol).

*e. What are the technical and/or policy challenges that you anticipate will need to be addressed for this idea to be successful?*

- Technical heterogeneity of DNA sequence data across multiple cohorts poses a challenge for data harmonization. However, this is a problem we have considerable experience in through our leadership of the gnomAD project, which has involved the assembly and joint variant calling of over 140,000 individuals derived from diverse sequencing cohorts.
- Accurate identification of true LoF variants, and distinguishing them from sequencing and annotation artifacts, is also challenging. We believe the only viable solution to this at the present time involves light automated filtering to capture common error modes, followed by deep manual curation of raw sequencing and annotation data.
- Accessing and harmonizing clinical data from multiple sources is a very significant obstacle to any project analyzing samples drawn from multiple cohorts. A new Health Data Research UK-funded proposal led by PI van Heel to aggregate clinical data from the East London Genes and Health project will contribute valuable infrastructure and expertise. In addition, the Broad Institute is leading efforts to aggregate clinical data from >1 million U.S. adults as a part of the AllofUs Precision Medicine Initiative, and these pipelines will be made available for the current proposal as well.
- Finally, we anticipate considerable regulatory obstacles associated with assessing consent, data use permissions, and the feasibility of genotype-based recontact across a diverse set of cohorts with widely varying consent protocols and regulatory environments. We will work with stakeholders of all included cohorts to establish a unified ethics protocol covering the work described in this proposal, and share this protocol with IHCC members to facilitate future cross-cohort collaboration.

# Toward a Federated Data Ecosystem

**Authors:** John Chambers, Ph.D., M.B.B.S. (Singapore SG100K study); Josh Denny, M.D., M.S. (Vanderbilt and *All of Us*); Mark Effingham, Ph.D., M.Sc. (UK Biobank); David Glazer (Verily Life Sciences and *All of Us*); Anthony Philippakis, M.D., Ph.D. (Broad Institute and *All of Us*).

## 1. Idea Summary

The life sciences are in the midst of a data revolution. Inexpensive and accurate genome sequencing is a reality, advanced imaging is routine, clinical data is increasingly stored in electronic form, and digital health wearables are poised for wide adoption. In principle, these innovations—and the massive data sets they produce—have brought us to the threshold of a new era in biomedicine, one where the data sciences hold the potential to propel our understanding and treatment of human disease.

In practice, we are stymied by our inability to overcome the challenge of federated data sharing. It is clear to all that we must be able to query data across multiple institutions, as no single effort has a sufficient N, especially for underrepresented populations, to achieve statistical saturation. Yet genomic and clinical data is generated in geographically disparate regions, subject to an array of regulatory environments, and analyzed with a variety of tools. No one entity can serve the needs of such a wide variety of stakeholders, and an array of interoperable solutions are needed.

Here, we propose a flagship scientific initiative aimed at reducing federated data sharing to practice. *Our plan is to bring together three cohorts—the UK Biobank, the All of Us Research Program, and the Singapore 100K Cohort—in a federated fashion*. Each dataset will be stored in its own data repository, and each group will maintain control over who can access its dataset. On top of these three repositories, we will instantiate a system of workspaces pointing to these data repositories, providing a secure environment for cross-analysis *without having to store multiple copies of each dataset.*

We believe this pilot will be an exemplar for others to emulate. By creating the extensible architecture described here, we hope to nucleate an ecosystem of like-minded, interoperable data environments.

## 2. Structured Abstract

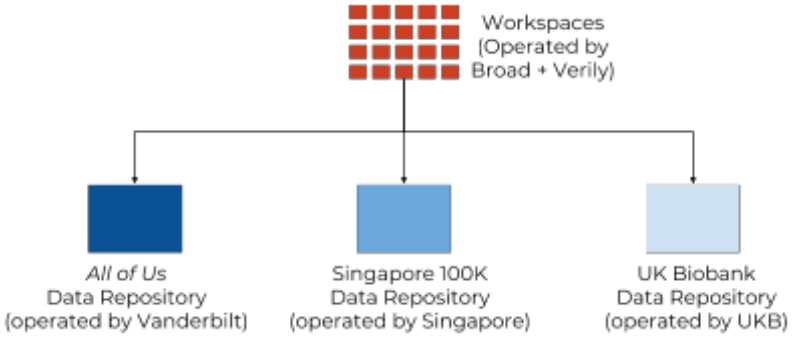
*A.* *Background and rationale*: The world's current model for sharing data is to place it on servers for researchers to download. This poses three challenges, however: 1) it is expensive, as it requires a copy of every dataset at every institution, 2) it is insecure, as it is difficult to track and audit who has touched a dataset and for what research purpose,
3) it limits access, as only researchers at well-resourced institutions have sufficient computational infrastructure.

Cloud technologies hold the promise of a new and more powerful model of data sharing. By storing data in an environment that co-locates compute and storage, we can analyze data in situ, rather than via downloading. By bringing researchers to the data, rather than data to the researchers, we can overcome the aforementioned challenges.

*B.* *Idea*: With this new opportunity, however, comes a new challenge. As cloud-based data environments emerge to support initiatives around the world, how will we share data across them?

One approach is to create a single, monolithic data platform. This model is conceptually simple and has precedents in the world of technology. It is the wrong answer for the life sciences, however, as it will not provide the flexibility to accommodate the wide range of scientific needs and the diverse regulatory requirements of different jurisdictions.

*Instead, our idea is to create a federated system that allows each cohort to maintain control of their data, while also allowing these cohorts to be easily cross-analyzed.*




Our basic architecture is composed of the following components:

a. A system of three data repositories holding the *All of Us*, UKB, and SG100K cohorts, respectively. Each repository will be operated by its own initiative.
b. A system of workspaces — secure environments where researchers can analyze data — that point to the data in each of these repositories. Broad and Verily will operate one such system, but others are free to participate as well.

C. *Impact*: This effort will be transformative in two ways:
   a. It will enable researchers to seamlessly perform analyses across these three cohorts, leading to discoveries in population health and human genetics.
   b. It presents an extensible model that can be extended to additional cohorts, holding the potential to nucleate a federated ecosystem of partners.

D. *Timeline:* We propose that this project can be completed in three years. Key Milestones are the following:
   a. Year 1:
      i. Deploy initial versions of the architectural components listed above in support of all three cohorts.
      ii. Partial harmonization of data models across the three cohorts, with a goal of further harmonization during the course of this award.
   b. Year 2:
      i. Address regulatory, security, and other data stewardship issues (including GDPR), so that these datasets can be cross analyzed.
      ii. The first group of researchers has gained access to do cross-analysis of multiple datasets as "beta testers."
   c. Year 3:
      i. Outreach to other cohort owners to ensure a generalizable model.
      ii. The system is available for use by researchers everywhere.
      iii. Execute a small series of demonstration projects that utilize data across the three resources.

E. *Estimated funding required:* $25m, spread over three years.

**3. Additional Information**

A. *Why does this idea require multiple large cohorts? Which cohorts would be required?* Our goal is

to reduce the concept of federated data sharing to practice. We have selected these cohorts (UK Biobank, *All of Us*, and Singapore SG100k) for the following reasons:

a. They are geographically disparate (US, Europe and Asia), and therefore subject to different organisational and regulatory considerations. This will force us to address these challenges at an early stage of development, rather than shy away from them.
b. "Three" represents a large enough collection of cohorts to be ambitious, yet also tractable over a three-year period. If we are successful, it will be clear to all that our model can be extended to an arbitrary number of cohorts.
c. Our groups know each other well, and we have already been discussing how we might work together more closely. Thus, we are confident that we are aligned around a common vision for federated data sharing.

We stress, however, that our intent is to incorporate additional cohorts into this framework, especially during Years 2 and 3 of the award. We are particularly enthusiastic about including groups from Africa, Latin America, and Europe to demonstrate how this might be a truly global model.

*B.* *What kinds of data/sample access will be required?*
We propose that the data repositories listed above will hold the entire de-identified datasets from these three cohorts.

Researchers will apply to each of these cohorts as they would today but, when they are granted access to the dataset (or a subset of it), they will then be able to analyze it in a secure, cloud-based environment. Moreover, assuming that the researcher has gained access to multiple cohorts, she will be able to cross-analyze them in thissecure,
cloud-based environment.

*C.* *What additional assays or data collection will be required?*
We do not propose to generate additional data for these three cohorts as part of this proposal.

*D.* *What is the data analysis plan?*
We plan to have a small number of demonstration project that will be developed in later years of the grant in consultation with the scientific community.

*E.* *What are the technical and/or policy challenges that you anticipate will need to be addressed for this idea to be successful?*
We anticipate working through the following challenges:

**Data Locality** - Many countries require that health data must remain in data centers physically located in that country. These restrictions are often complex and unclear, however. Consider the following:

a. Does the requirement for geographic locality apply to data that has already been stripped of obvious personal identifiers?
b. Does geographic locality apply to aggregate-level results? If so, then it would be impossible to disseminate scientific findings via publication that are printed abroad. If not, then how aggregated is aggregated-enough? Does aggregating across 2 individuals count?
c. If a researcher is traveling to another country, is she prohibited from doing research by logging into a datacenter in her home country, as individual-level data will appear on her laptop screen while she is abroad?

A key goal of this award will be to work with regulators in each of these three countries to

understand the scope of the possible, with a strong eye towards finding ways of creating one or more “data Switzerlands” — i.e., a secure enclave where it is possible to cross-analyze datasets from different countries of origin, while maintaining regulatory compliance.

2. **Researcher Identities** - For many consumer technology services, such as Spotify or Twitter, the user does not need to disclose her true identity and can instead use a pseudonym. By contrast, we require researchers to declare their true identity when requesting access to datasets.

   While this is clearly appropriate and needed for research oversight, the scientific community has not yet sufficiently addressed two technological challenges: 1) Who are the providers of researcher identities? and 2) What are the systems for ensuring that a researcher is who she says she is (i.e., mechanisms of “identify proofing”). While some systems exist, such as eRA Commons in the United States and EGA in Europe, they are not necessarily adopted or understood around the world.

   Achieving federated data sharing will require addressing these challenges around researcher identity. We propose to allow each cohort to continue to use its own system of researcher identities, but generate an additional layer of security protocols that link identities across these three repositories. We will also establish more rigorous methods of identity-proofing researchers. Through this, we hope to create a system that is federated, yet gives the experience of being a “single sign-on” for researchers.

3. **Data Use and Researcher Conduct** - Human subjects data is notable for having two axes of access control — one based on *who* the researcher is (i.e., is this researcher authorized to see this dataset?), and the other based on *what* the researcher is doing (i.e., this dataset was consented for only noncommercial diabetes research — is the researcher’s purpose consistent with this restriction?). Thus, a given researcher can only access a dataset if (i) they have permission to do so, and (ii) their research purpose is consistent with the secondary data use restrictions as expressed in the informed consent form. While permissionsare

   made machine readable through commonplace authorization protocols, we lack a machine-readable framework for expressing data use restrictions.

   A key goal of this effort will be to instantiate machine-readable ways of expressing data use and standardize them across these efforts. Also, we will establish a common code of conduct for researchers that clearly articulates inappropriate secondary uses of data (e.g., researchers should not try to
   re-identify research participants).

   Dr. Philippakis is a leader in this domain, having previously chaired the GA4GH workstrop on Data Use and Researcher Identity, and also having been the primary author of the *All of Us* Data Access Framework.

4. **Computational Scalability** - The computational challenges associated with storing, sharing, and analyzing these three massive datasets should not be underestimated. During the lifetime of this award, we anticipate that there will be more than 1m whole genomes available across these cohorts, along with other massive data types such as imaging and sensors.

   Each of our groups has extensive experience at building scalable cloud-based solutions. For example, Dr. Philippakis’ team at Broad currently processes more than 30 TB of data

per day, and has more than 100 PB of data under management. Similarly, as part of Alphabet, Verily Life Sciences has particular expertise in distributed systems and cloud technologies that we will leverage to achieve the goals proposed here.

5. **Development of Standards** - The Internet and the World Wide Web would not have been created without the TCP/IP and HTML standards, respectively. In a parallel way, achieving a federated data sharing ecosystem will require the creation and adoption of international standards around life science datasets.

   Members of our team are leaders in standard-setting. For example, Dr. Philippakis was the co-chair of the GA4GH Data Use and Researcher Identity workstream. Similarly, Mr. Glazer continues to co-chair the Cloud workstream of GA4GH. We will leverage this strong track-record to ensure that the system built here is well-aligned with emerging standards for genomic and clinical data.

# Trial of Polypill and Healthy Lifestyle Across Several Cohorts within IHCC Consortium to Prevent Premature CV Disease Morbidity and Mortality and Define a More Accurate CVD Risk Score

**Author:** Reza Malekzadeh, M.D., Professor of Medicine, Digestive Disease Research Institute, Tehran University of Medical Sciences, Shariati Hospital, Tehran, Iran

One of the major health issues at present time both in developed and specially the developing countries is Premature and preventable death and disability (PPDD)[1]. PPDD is one of the major of Sustainable development goals which all countries are invited to pay attention for achieving it by year 2030[2].Based on our experience in Golestan Cohort study (GCS) 63% of death are premature (<70) and 30% are very premature (>60). Based on GBD 2017 study about 50% of death at global level are premature[3-6]. We now have adequate and good quality evidence that at least 40% of premature death and disability are preventable specially the cardiovascular mortality using life style and simple pharmacological preventive measures[7-8]. The past 20 years has seen impressive advances in cardiovascular (CV) care and decreases in patient mortality from CV disease. Despite this impressive gain, there is still substantial room for further progresses in CV care.

A recent study looking at the interplay between lifestyle and genetic risk factors for CV disease in one cross-sectional study and three prospective cohorts found that while high genetic risk (hazard ratio, 1.91) independent of healthy lifestyle result in increased risk of CV disease but within any genetic risk category, adherence to a healthy lifestyle was associated with a significantly decreased risk of CV disease[8].

The PolyIran pragmatic cluster randomized controlled trial nested in GCS[9] have shown that using polypill and healthy lifestyle for 5 years was effective in reducing the CV disease morbidity and mortality by 50%.

We would like to propose a trial of polypill and healthy lifestyle across several cohort within IHCC consortium to prevent premature CV disease morbidity and mortality .This trial In addition to saving life and improving the CV health could also create a good opportunity for IHCC scientists to better define a CV risk score which combine genetic risk alleles and lifestyle factors factor which would become the best risk score[10] which need intervention for prevention of CV disease for future[9-10] .

## References

1. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017.Lancet. 2018 Nov 10;392(10159):1789-1858.

2. GBD 2017 SDG Collaborators.Measuring progress from 1990 to **2017** and projecting attainment to 2030 of the health-related Sustainable Development Goals for 195 countries and territories: a systematic analysis for the Global Burden of Disease Study **2017**.Lancet. 2018 Nov 10;392(10159):2091-2138.

3. Pourshams A, Khademi H, Malekshah AF, Islami F, Nouraei M, Sadjadi AR, Jafari E, Rakhshani N, Salahi R, Semnani S, Kamangar F, Abnet CC, Ponder B, Day N, Dawsey SM, Boffetta P, Malekzadeh R.**Cohort** Profile: The **Golestan Cohort** Study—a prospective study of oesophageal cancer in northern Iran.Int J Epidemiol. 2010 Feb;39(1):52-9.

4. Nalini M, Oranuba E, Poustchi H, Sepanlou SG, Pourshams A, Khoshnia M, Gharavi A, Dawsey SM, Abnet CC, Boffetta P, Brennan P, Sotoudeh M, Nikmanesh A, Merat S, Etemadi A, Shakeri R,

Sohrabpour AA, Nasseri-Moghaddam S, Kamangar F, Malekzadeh R.Causes of premature death and their associated risk factors in the Golestan Cohort Study, Iran. BMJ Open. 2018 Jul 18;8(7):e021479.

5. Sepanlou SG, Sharafkhah M, Poustchi H, Malekzadeh MM, Etemadi A, Khademi H, Islami F, Pourshams A, Pharoah PD, Abnet CC, Brennan P, Boffetta P, Dawsey SM, Esteghamati A, Kamangar F, Malekzadeh R. Hypertension and mortality in the Golestan Cohort Study: A prospective study of 50 000 adults in Iran. J Hum Hypertens. 2016 Apr;30(4):260-7.

6. Nalini M, Sepanlou SG, Pourshams A, Poustchi H, Sharafkhah M, Bahrami H, Kamangar F, Malekzadeh R. Drug Use for Secondary Prevention of Cardiovascular Diseases in Golestan, Iran: Results From the Golestan Cohort Study.Arch Iran Med. 2018 Mar 1;21(3):86-94.

7. Nugent R, Bertram MY, Jan S, Niessen LW, Sassi F, Jamison DT, Pier EG, Beaglehole R.Investing in non-communicable disease **prevention** and management to advance the Sustainable Development Goals.**Lancet**. 2018 May 19;391(10134):2029-2035.

8. Chêne G.**Prevention** of the causes of **premature** illness and **death**: making it happen.**Lancet Public Health**. 2017 Feb;2(2):e69-e70.

9. Khera AV, Emdin CA, Drake I, Natarajan P, Bick AG, Cook NR, Chasman DI, Baber U, Mehran R, Rader DJ, Fuster V, Boerwinkle E, Melander O, Orho-Melander M, Ridker PM, Kathiresan S.**Genetic Risk**, **Adherence** to a **Healthy Lifestyle**, and **Coronary Disease**.N Engl J Med. 2016 Dec 15;375(24):2349-2358.

10. Fahimfar N, Khalili D, Sepanlou SG, Malekzadeh R, Azizi F, Mansournia MA, Roohafza H, Emamian MH, Hadaegh F, Poustchi H, Mansourian M, Hashemi H, Sharafkhah M, Pourshams A, Farzadfar F, Steyerberg EW, Fotouhi A.Cardiovascular mortality in a Western Asian country: results from the Iran **Cohort** Consortium. BMJ Open. 2018 Jul 5;8(7):e020303.

11. Ostovaneh MR, Poustchi H, Hemming K, Marjani H, Pourshams A, Nateghi A, Majed M, Navabakhsh B, Khoshnia M, Jaafari E, Mohammadifard N, Malekzadeh F, Merat S, Sadeghi M, Naemi M, Etemadi A, Thomas GN, Sarrafzadegan N, Cheng KK, Marshall T, Malekzadeh R.Polypill for the prevention of cardiovascular disease (PolyIran): study design and rationale for a pragmatic cluster randomized controlled trial. Eur J Prev Cardiol. 2015 Dec;22(12):1609-17.

12. Merat S, Poustchi H, Hemming K, Jafari E, Radmard AR, Nateghi A, Shiravi Khuzani A, Khoshnia M, Marshall T, Malekzadeh R. PolyPill for Prevention of Cardiovascular Disease in an Urban Iranian Population with Special Focus on Nonalcoholic Steatohepatitis: A Pragmatic Randomized Controlled Trial within a **Cohort** (PolyIran - Liver) - Study Protocol. Arch Iran Med. 2015 Aug;18(8):515-23.

# Using the Norwegian Mother and Child Cohort Study (MoBa) to Explore Disease Etiology, Genotypic Fitness and Health of Adolescents

**Principal Investigator:** Per Magnus, M.D., Ph.D.

## *1.* Idea Summary

Many countries around the world are now seeing low fertility rates. Increasing age at childbirth is likely to play an important role. However, there are indications of biological changes, such as declining semen quality and increasing difficulties for couples to conceive. Exposure to environmental contaminants in fetal life may influence later fecundity. MoBa is a nation-wide cohort of trios (mother, father, child) with more than 114,000 children. The oldest children are now 19 years old. The recruitment was during their mothers' pregnancies in the years 1999-2008. We now propose to examine 5,000 male and 5,000 female participants, aged 18 to 20 years, at fertility clinics during the years 2020 to 2022. Year 2020 will be a pilot year. MoBa has collected biomaterials from mothers, fathers and children from pregnancy and birth (cord blood for the children). Extensive questionnaires during their mothers' pregnancies and their own childhood have provided information on exposures and life styles (see www.fhi.no/moba). Genome-wide SNP genotyping has been performed on about 30,000 trios in MoBa, and will be complete for more than 50,000 trios during the coming year. Genome-wide methylation and exome-sequencing have been performed in some trios. Including the father has opened for detecting de novo mutations, and the generational design opens for studies of relative fitness of alleles, genotypes and phenotypes. At the clinics, in addition to anthropometrics and a state-of-the-art fertility examination, we will collect new blood samples for measures of biomarkers (exposures and early signs of disease) as well as repeated methylation profiles and other omics. In addition, participants will respond to short (5-minute) mobile phone questionnaires every other month – covering a series of items over time. Due to the unique personal identification numbers, linkage to health registries will be performed. Adolescence – 15 to 25 years – is an understudied period of transitions determining later health. Our trio design opens for a series of exciting research options – reduced fecundity being only one of them.

## *2.* Structured Abstract

Background and rationale: The Norwegian Mother and Child Cohort Study (MoBa)[1-3] was initiated and planned in parallel with the Danish National Birth Cohort (DNBC)[4]. Due to differences in funding, DNBC started recruitment of pregnant women in 1996, while MoBa started in 1999. The oldest children are now in their early twenties or late teens. The main idea behind both cohorts was to test hypotheses about early life causes of adult disease.

For all countries, non-communicable diseases (NCDs) pose major challenges. Adolescence is of central importance for understanding the development of NCDs and mental health[5]. In this period people choose educations and occupations, they test out and establish new life styles and develop independent social networks. Choices and habits established through adolescence is important for the risk of later cardiovascular diseases, and the maintenance or development of obesity. Major mental disorders as well as some neurological and autoimmune disorders have their debut during this age period. In Norway and Denmark, the cohorts will link overt diseases to the databases through health registries.

However, some health-related traits, signs and symptoms can only be included into the cohorts through dedicated clinical examinations. One example is fecundity, the biological ability to have children. The main reason for the declining fertility rates is postponement of childbirth, but there is a growing concern that particularly male infertility is increasing. There is an increase in the occurrence of cancer in young adults[6], including a strong increase in testicular cancer In Norway and other Nordic countries[7]. A recent meta-analysis shows falling sperm counts[8]. There are speculations that these trends are due to effects

of toxicants[9], even exposure as early as fetal life may play a role. Female fecundity is more difficult to measure and less studied. We know little about time trends in female fecundity.

Idea: MoBa was set up to respond to hypotheses and speculations about effects of early exposures on later health and disease. A life-course approach requires repeated sampling of exposures and biomarkers from fetal life through adolescence and into adulthood. Our idea is to follow all MoBa participants, who continue to give consent and to invite 10,000 participants for clinical examination, with a plan to repeat the examinations at 5-year intervals.

Impact: Politicians are struggling to find ways to respond to issues such as the obesity epidemic, the falling fertility rates, the prevention of NCDs and mental illnesses. MoBa-based research will provide them with policy options. A full exploitation of omics-methodology will fertilize research into etiology and mechanisms, including biological aging and transgenerational transmission of health. The comparison of gene frequencies across generations in trios is the basis for evolutionary genetics, including allele-specific fitness and mutation rates. The social mechanisms influencing health, fertility and function in early adulthood will be studied in parallel.

Timeline: In September 2019, the oldest MoBa-child will be 20 years old. We will use 2020 as a pilot year involving only one fertility clinic and a few hundred participants, and expand to full recruitment in three clinics during 2021-22.

Estimated funding: The clinical set up, the sampling, processing, storing and analyses of biomaterials for 10,000 participants is expensive. The examinations will be performed at university clinics in Oslo, Bergen and Trondheim, the three largest cities of Norway. The needed funding for data collections and clinical examinations is 2.9 million USD. The omics analyses will cost 4.9 million USD. The total cost is 7.8 million USD.

## *3.* Additional information

Other cohorts: This proposal has been discussed with the leaders of the DNBC. They are enthusiastic on collaborating in a similar follow-up of the adolescents and young adults in both cohorts. We will strongly encourage parallel investigations for replications and increased sample size. The overlap in inclusion periods, but also some discrepancies enable us to stretch the period of fetal exposure by using both cohorts. In the DNBC, sperm collections and measures of maturation have commenced[10]. We are not aware of other large birth cohorts with adolescent participants that have banked biomaterials from fetal life.

Data/sample access: MoBa is an open resource for bona fide researchers from all over the world, in accordance with the original consent from the parents and the new consent to be collected from the children. In addition, approvals from Norwegian ethics committees and relevant data security measures are required.

Additional assays and data collections:
We will invite MoBa participants and continue recruitment to clinical and fertility examination until we have recruited 5,000 young women and 5,000 young men who are among the 114,000 participating "children" in the MoBa cohort. Clinical data and blood samples, as well as frequent electronic questionnaire will be collected. We have the infrastructure at our institute for handling and storing data through questionnaires, clinical examinations and sampling of biomaterials.[1-3] The bio-samples will be processed and stored at the Norwegian Institute of Public Health. DNA extracted from fresh blood and will be analyzed for genomics.

Data collections will consist of clinical examinations at fertility clinics in Norway. The leader of the organization of these clinics is affiliated with our Centre of Excellence and will supervise the clinical

exams. The other part of the data-collection is the frequent mobile phone-based questionnaires, which will be sent out to all participants regularly. These questionnaires will be short and cover different topics in each round.

Data analysis plan: The project will be lead and coordinated by Per Magnus. The MoBa staff and senior researchers at the Centre of Excellence and the Department of Genetics at NIPH will take responsibility for data-collections, bio-banking and the analytic work. The involved research groups have broad experience and are experts in utilizing trio data. We have ongoing projects within the fields of biological maturation and ageing starting in fetal life (for instance the epigenetic clock and telomeres), and have established groups for fertility research, including consequences of assisted reproductive technologies.

We will employ different methodologies:

*Approach 1: Epidemiological analyses*
The detailed information in MoBa will enable us to study the interplay between genes, fetal exposures, childhood environments and lifestyles on the one hand, and adolescent health and fecundity measures on the other. We will use traditional epidemiological approaches, and combine these with new methods for causal inference.

*Approach 2: Use family designs and generational data, including genetic and epigenetic analyses*
In registries and MoBa there is multigenerational data on families. We will use several designs to enable control for unmeasured factors that may influence associations. Researchers actively involved in the project have developed analytical methods and software using trio designs (mother-father-child).[11,12] This methodology is specially designed to study effects of fetal genes, maternal genes, parent-of-origin, and fetal-maternal gene interactions. In addition, it allows evaluation of gene-environment interactions, including maternal environmental exposures, as well as gene-methylation interactions. We will expand analytical approaches where we combine epigenetic and genetic data from trios.

*Approach 3: Use genes as instrumental variables*
As observational studies based on registries and cohorts are prone to bias, we will add approaches where we can make stronger causal interpretations.[13] Genetic variants can be used as instrumental variables in Mendelian Randomization (MR) analyses,[14] a design which mimic randomized trials. We will use known genetic determinants for epigenetic age acceleration and telomeric length as instruments when evaluating biological aging against measures of fecundity [15], and we will use known genetic determinants of lifestyle characteristics (e.g. education, smoking and BMI) when evaluating their effect on health outcomes. Recent genome wide association studies (GWAS) have identified a number of SNPs associated to fertility[16-18] which will be related to the fecundity variables.

We will use our expertise in perinatal and genetic epidemiology to select suitable models. New advances in measuring DNA-methylation (e.g. the Illumina EPIC array) will require development and adaption of existing age prediction models. We will be supported by qualified expertise in Norway, U.K., and the U.S., who are at the forefront in the development of suitable models, taking into consideration the requirements for sufficient statistical power.[19,20]

Project group and resources:

This project will be based within a multidisciplinary collaborative environment, facilitated by the NIPH, the MoBa team, and the Centre of Excellence in Fertility and Health (www.cefh.no ). This project is divided into three work packages (WPs). The WPs will collaborate closely. Each WP is responsible for defined main aims. The PI of MoBa, Per Magnus, will lead the WPs and ensure that tasks and deliveries are reached within the planned timeline.

*WP0: Organization and logistics*

*WP1: Clinical examination of adolescents and data collection through questionnaires during adolescence*
*WP2: Epidemiologic analysis, 'omics and adolescent health*

The research team: The project will involve key senior researcher and international experts within perinatal epidemiology, genetics/epigenetics, statistics and medicine:

- Per Magnus, NIPH, Director of Center for Fertility and Health (CeFH), PI of MoBa, expert in epidemiology and medical genetics
- Siri Håberg, NIPH, Deputy director of CeFH, expert in perinatal epidemiology
- Håkon Gjessing, NIPH, Head statistician, expert in statistical genetics
- Jon Bohlin, NIPH, Senior researcher, expert on epigenetic clock
- Astanand Jugessur, NIPH, Senior researcher, expert in genetic epidemiology
- Debbie Lawlor, Professor of Epidemiology, University of Bristol, and Deputy Director, UK Medical Research Council Integrative Epidemiology Unit, U.K., expert on genetic instruments
- Allen J. Wilcox, Head of Reproductive Epidemiology at the National Institute of Environmental Health Sciences (NIEHS), USA, expert in perinatal epidemiology
- Stephanie J. London, founder and PI of the PACE consortium, expert in environmental exposures, epigenetics and health (NIEHS)
- Abraham Aviv, Rutgers University, New Jersey, expert in telomere research
- Junior researchers: We have several PhDs and Post Docs that will actively participate in the work packages. We will apply for additional funding to recruit more young researchers within each of the analytic approaches.

The clinical examination team will include:

- Ketil Størdal, Head of the pediatric organization in Norway
- Petur Juliusson, Expert pediatrician within growth and adiposity
- Pål Suren, Pediatric specialist in neurodevelopmental disorders
- Hans Ivar Hanevik, Clinician and head of the Norwegian Organization of Assisted Conceptions
- Liv Bente Romundstad, Gynecologist, head of fertility clinic.

Technical and political challenges:
The project will increase the understanding of fetal origins of health and diseases. It will describe and analyze fecundity and increase fertility awareness. This will support young people planning their lives, and give policymakers better background for actions that may enhance population fertility. We do not foresee political problems with this research. On the contrary, the recent fertility decline in Norway has raised concerns and public debate. By giving access to data for researchers from all over the world, it will establish MoBa as a playground for basic and applied research into a large array of health issues, in line with the Norwegian Research Council's efforts to increase internationalization of research. Most of the infrastructure necessary for this project is already in place. Funding is required to secure personnel for establishing data collections and clinical examinations, and detailed piloting during the first year is needed for efficient data collection. We will give young people fertility counselling in a professional

setting within an established cohort. The pilot phase and careful start of the project is necessary to ensure a sufficient participation rate.

**References:**


1. Magnus P, Birke C, Vejrup K, et al. Cohort Profile Update: The Norwegian Mother and Child Cohort Study (MoBa). *International journal of epidemiology.* 2016;45(2):382-388.
2. Ronningen KS, Paltiel L, Meltzer HM, et al. The biobank of the Norwegian Mother and Child Cohort Study: a resource for the next 100 years. *Eur J Epidemiol.* 2006;21(8):619-625.
3. Magnus P, Irgens LM, Haug K, Nystad W, Skjaerven R, Stoltenberg C. Cohort profile: the Norwegian Mother and Child Cohort Study (MoBa). *International journal of epidemiology.* 2006;35(5):1146-1150.
4. Olsen J, Melbye M, Olsen SF, et al. The Danish National Birth Cohort—its background, structure and aim. *Scandinavian journal of public health.* 2001;29(4):300-307.
*5.* Patton GC, Olsson CA, Skirbekk V, et al. Adolescence and the next generation. *Nature.* 2018;554:458.
6. Burkhamer J, Kriebel D, Clapp R. The increasing toll of adolescent cancer incidence in the US. *PloS one.* 2017;12(2):e0172986.
7. Ylonen O, Jyrkkio S, Pukkala E, Syvanen K, Bostrom PJ. Time trends and occupational variation in the incidence of testicular cancer in the Nordic countries. *BJU international.* 2018;122(3):384-393.
8. Levine H, Jorgensen N, Martino-Andrade A, et al. Temporal trends in sperm count: a systematic review and meta-regression analysis. *Human reproduction update.* 2017;23(6):646-659.
9. Minguez-Alarcon L, Williams PL, Chiu YH, et al. Secular trends in semen parameters among men attending a fertility center between 2000 and 2017: Identifying potential predictors. *Environment international.* 2018;121(Pt 2):1297-1303.
10. Brix N, Ernst A, Lauridsen LLB, et al. Timing of puberty in boys and girls: A population-based study. *Paediatric and perinatal epidemiology.* 2019;33(1):70-78.
11. Gjessing HK, Lie RT. Case-parent triads: estimating single- and double-dose effects of fetal and maternal disease gene haplotypes. *Annals of human genetics.* 2006;70(Pt 3):382-396.
12. Gjessing HK. HAPLIN. 2018; (folk.uib.no/gjessing/genetics/software/haplin) 2018.
13. Smith GD, Lawlor DA, Harbord R, Timpson N, Day I, Ebrahim S. Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. *PLoS medicine.* 2007;4(12):e352.
14. Katikireddi SV, Green MJ, Taylor AE, Davey Smith G, Munafo MR. Assessing causal relationships using genetic proxies for exposures: an introduction to Mendelian randomization. *Addiction (Abingdon, England).* 2018;113(4):764-774.
15. Lu AT, Xue L, Salfati EL, et al. GWAS of epigenetic aging rates in blood reveals a critical role for TERT. *Nature communications.* 2018;9(1):387.
16. Lunetta KL, Day FR, Sulem P, et al. Rare coding variants and X-linked loci associated with age at menarche. *Nature communications.* 2015;6:7756.
17. Perry JR, Corre T, Esko T, et al. A genome-wide association study of early menopause and the combined impact of identified variants. *Human molecular genetics.* 2013;22(7):1465-1472.
18. Barban N, Jansen R, de Vlaming R, et al. Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nature genetics.* 2016;48(12):1462-1472.
19. Tsai PC, Bell JT. Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation. *International journal of epidemiology.* 2015;44(4):1429-1441.
20. McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews Genetics.* 2008;9(5):356-369.

# Whole Genome Sequencing for Phenome-Wide Association Studies

**Author:** Francine Grodstein, Sc.D.

## 1. Idea Summary

Whole genome sequencing (WGS) across the diverse global populations of IHCC will allow for an in-depth investigation of the role of rare variants in human health and disease. WGS will allow for broad phenome-wide association studies (PheWAS) of rare variants, as well enhanced development of risk prediction models encompassing the broadly diverse populations.

## 2. Structured Abstract

Background and rationale: While GWAS studies have contributed a lot of knowledge about the impact of common variants on risks of chronic diseases, whole genome sequencing (WGS) allows for a deeper look at rare variants, which likely contribute to risk of disease as well as allow for better understanding of the underlying causal mechanism. Phenome-wide association studies (PheWAS) of rare variants, leveraging the very large sample size available across the IHCC. In light of the role of genetic risk in risk prediction modeling, harmonizing germline genetic data across cohorts will allow for calculation of absolute risks that currently we can only estimate using sampling schema from case-control studies. Further, incorporating genetic information from global population samples will allow for better calibration of risk prediction models across different populations.

Idea: We propose to conduct WGS in the IHCC. By leveraging archived biospecimens available across cohorts in the IHCC, across multiple countries and continents, we can investigate differences in both rare and common variants by race/ethnicity. We can also examine how phenotypes, such as obesity and cardiometabolic disease, relate to rare variants. Finally, as follow-up after blood collections continues in cohorts, we can begin to compare these omics profiles in incident cases versus controls selected on the basis of specific chronic diseases (i.e., cancers, cardiovascular diseases, cerebrovascular diseases). Alternatively, we could limit the investigation to focus on one disease, such as diabetes, that would allow for cross-sectional investigations as well as an initial nested case-control study as a pilot project.

Impact: WGS will allow an in-depth look at rare variants across the diverse populations of the IHCC, and their role in phenotypes, health and disease. Leveraging the strength of numbers in the IHCC, we will be able to study in detail rarer diseases, such as pancreatic cancer or amyotrophic lateral sclerosis (ALS).

Timeline: The work involved in this proposed idea would include: 1) WGS assay in archived biospecimens with DNA within each participating cohort; 2) detailed analysis plans to be carried out within each cohort, including PheWAS and genetic risk prediction models; 3) meta-analysis of results across cohorts. We anticipate the WGS would be completed in years 1 and 2, with concurrent development of analytic plans. Analysis of individual cohort data would be conducted in year 2 and the start of year 3, with meta-analysis across cohorts conducted in year 3.

Estimated funding required: WGS of extracted DNA would be required, at a cost of approximately $350 per sample. Funding required for individual cohort needs would vary depending on infrastructure and cohort size, but would need to cover accessing and shipping blood samples, data management to process WGS data and incorporate with other cohort data, and analytic programming to conduct analyses. Additional funding may be required to cover the time of investigators coordinating the efforts within individual cohorts as well as investigator time to carry out this consortium project.

3. **Additional Information**

Although GWAS has been done broadly in several populations, WGS is not as widespread, and the incorporation of cohorts that capture ethnic/racial diversity as well as geographical and environmental diversity would allow for a large contribution to the field.

Extracted DNA will be required from participating cohorts. Required data would include lifestyle and disease factors of interest (e.g., smoking status and history, BMI (at points across the life course if available), disease history), as well as incidence of diseases of interest.

We assume most cohorts do not have broad WGS completed and will need to be completed de novo for this project.

To achieve the near-term accomplishment, we anticipate utilizing a meta-analysis approach for this project, designing analyses to be carried out within individual cohorts, and meta-analyzing across cohorts. Longer-term approaches could include incorporation and harmonization of data across cohorts to conduct pooled analyses.