International 100K Cohort Consortium

Meeting Summary

Grand Hotel
Gullteigur Conference Room, Reykjavik, Iceland

April 23-24, 2019

Hosted by the Global Genomic Medicine Collaborative (G2MC)

Vision for success:

A global platform for translational research (cohort to bedside and cohort to bench), informing biological/genetic basis for disease and impact on clinical care and population health

Global Genomic
Medicine Collaborative

International 100K Cohort Consortium

EXECUTIVE SUMMARY

In 2015, the National Institutes of Health (NIH) launched an effort to identify all large-scale prospective cohort studies involving at least 100,000 (100K) participants to explore the potential of bringing them together to address scientific questions none could answer alone. The goal was, and remains, to create a global network for translational research that will utilize large cohorts to enhance understanding of the biological, environmental, and genetic basis of disease and to improve clinical care and population health. This effort was discussed at the June 2016 Heads of International Research Organizations (HIROs) meeting and reached an agreement to commission the Global Genomic Medicine Collaborative (G2MC), working with the Global Alliance for Genomics and Health (GA4GH), to bring the cohorts together for an initial summit, which was held in March 2018 in Durham, North Carolina. It involved 61 cohorts from 33 countries, representing current and expected cohort sizes of 30 million and 35 million individuals, respectively.

Cohorts were invited based on four criteria: having 100K participants or more, were not selected for disease, having available biospecimens, and having at least the potential for longitudinal follow-up of participants. It soon became clear, however, that regions such as Africa, the Middle East, South America and South Asia were under-represented in such a compilation, so these criteria were relaxed somewhat to facilitate their participation. The outcome of the summit and initial plans for an International Hundred Thousand Plus Cohorts Consortium (IHCC) were discussed at the June 2018 HIROs meeting. They reached an agreement to support a broader effort to develop IHCC and hold a second summit in spring 2019. Cohort representatives were invited to volunteer for three teams focused on data standards and interoperability, scientific strategy and cohort enhancements, and policy and data sharing.

On April 23-24, 2019, at the Grand Hotel in Reykjavik, Iceland, G2MC hosted the second International Cohorts Summit. The objectives of this meeting were to identify scientifically meritorious cross-cohort research projects and collaborators willing to participate in them and to develop a scientific agenda for IHCC to bring to funding bodies. Prior to the Summit, IHCC solicited a Request for Ideas (RFI) to review and discuss during the second summit.

Approximately 117 representatives from 59 cohorts and 29 countries attended. Twenty-three proposals from the RFI solicitation were included for discussion at the meeting. The proposals encompassed several key themes, including data platforms, -omics studies, rare variant detection and human knock-outs, diversity and ethnicity gaps, and polygenic risk scores.

The Summit was organized into nine sessions so Teams could discuss issues during their breakout working sessions for subsequent reports and to identify next steps.

Five sessions were held on the first day of the conference: Session 1 provided the goals for the meeting and set the stage for the value and challenges of combining large cohorts and the opportunities for translational impact for health. Session 2 discussed opportunities for collaboration across broad geographies. Session 3 addressed data standards and infrastructure in accordance with the Data Standards and Infrastructure Team focus. Session 4 provided an overview of IHCC's scientific strategy and enhancements of our cohorts per the focus of the Scientific Strategy and Cohorts Enhancement Team. Session 5 discussed current policy and data sharing challenges per the focus of the Policy and Bio-Data Sharing Team.

Four sessions were held on day two, which started with a keynote address about big data and opportunities in healthcare, followed by a charge to the breakout groups. Session 6 hosted the breakout sessions for each of the working group teams. The team leads then gathered during Session 7 for the Team Lead synthesis in preparation for Session 8, which provided reports from the breakout sessions and discussion of synergies. Session 9 concluded the Summit with a summary of progress and action plans from the meeting.

Each of the teams progressed toward their goals during the summit. The Data Standards and Infrastructure Team identified two project proposals to establish data infrastructure and interoperability standards, starting broadly with a Cohort Discovery Atlas, then narrowing focus with a driver project involving a few large cohorts from discovery to federated analysis. The Scientific Strategy and Cohorts Enhancement Team focused the scientific agenda on projects that could demonstrate the cohesiveness of IHCC and the value of the collective group. They acknowledged a variety of possible projects that might use existing data, some of which could be low-resource, quick deliverables for IHCC, as well as possible projects that would require generating new data. The Policy and Bio-Data Sharing Team refocused their primary goal from purely data sharing to generating *collaboration among cohorts*, which might enable data sharing. They will develop a “first principles” document that would define the value of participation along with guidance for fair collaboration.

In addition to the team progress, the meeting revealed considerable interest in becoming a more formal consortium with a charter, guiding principles, membership guidelines, etc. Meeting attendees emphasized the need to engage under-represented populations and cohorts in low and middle income countries (LMICs) to advance the global scientific agenda.

The next steps and outcomes from this Summit include

- Develop and distribute a summary of the meeting
- Develop exemplar proposals for presentation to HIROs from each IHCC Team (Data Standards, Scientific Strategy, and Policy/Bio-Data Sharing)
- Bring priorities to the HIROs
- Develop a charter and guiding principles for the consortium, potentially including a more formalized sign-up process and consideration of the expectation that Team co-leads cover the breadth of time zones to promote inclusivity
- Conduct a second solicitation of RFIs with sufficient advance notice, time to respond and proposal reviews by a broader section of the consortium
- Promote opportunities for trainees and junior investigators to participate in IHCC activities
- Circulate the list of cohorts in attendance and those that IHCC has knowledge of to ensure that no cohorts have been overlooked
- Based on the teams’ and proposed projects’ progress, host a third summit in 2020 that includes as many cohorts represented as possible and industry participation. This summit will present progress on the ongoing projects.

Global Genomic Medicine Collaborative

International 100K Cohort Consortium

INTERNATIONAL COHORTS SUMMIT 2

Hosted by the Global Genomic Medicine Collaborative (G2MC)

Grand Hotel, Gullteigur Conference Room, Reykjavik, Iceland
April 23-24, 2019

MINUTES

Day 1: April 23, 2019, 8:30 AM – 7:10 PM

SESSION 1 – INTRODUCTION AND BACKGROUND
CHAIR: GEOFFREY GINSBURG
8:30 – 9:45 AM

Welcome and Introductions – Geoffrey Ginsburg *(Duke University, G2MC, USA)*
Geoff Ginsburg opened the meeting by welcoming the group to Iceland for the second meeting of the International One-Hundred Cohorts Consortium (IHCC). He acknowledged the efforts of IHCC co-chairs (Teri Manolio, Peter Goodhand), the Executive Director of G2MC (Teji Rakhra-Burris), and Palladian Partners (Ida Donner and colleagues) in organizing the event. He also referenced the meeting resources of the e-Programme including Request for Ideas (RFIs) submissions, IHCC website (ihcc.g2mc.org), and a live stream of the meeting available at https://ihcc.g2mc.org/ics2019.

Welcoming Remarks – Guðni Thorlacius Jóhannesson *(President of Iceland)*
President Jóhannesson highlighted the collaborative environment of the meeting, with representatives from 29 countries in attendance. He acknowledged the unique genetic resources of Iceland, with detailed genealogy records that allow Icelanders to draw connections between themselves and distant relatives. He challenged the meeting attendees to follow ambition and self-belief, coupled with modesty and a willingness to listen to others, as the key to progress in making the world a better place.

Using Big Data as a Role Model to Improve Population Health in Iceland – Birgir Jakobsson *(Medical Advisor to the Minister of Health, Iceland)*
Dr. Jakobsson discussed the potential of Big Data to revolutionize healthcare and the challenges associated with implementing the knowledge gained from it. In Iceland, there are many well-kept data sources that enable collaboration coupled with a low expectation of personal integrity that could allow big data to be used for population health improvement. Some healthcare providers have started to use their data to predict patient outcomes or utilize genetic information to employ precision medicine for their patients. Soon after starting as the surgeon general, he attended a global seminary on big data and they created a working group on how to utilize big data in Iceland and create a national strategy with stakeholder buy-in. In order to drive value in the health system and improve population health, they needed to prioritize diseases of focus, collect and curate the data for predictive analysis, and eventually, communicate the results to providers and the public. The project would be a continuous improvement without an end-point. This strategy is now being considered in parliament with hopes for a five-year implementation plan. This structure in Iceland can be used as a model of what may be possible in other environments.

Cohorts: A Collective Vision – Mary De Silva *(Wellcome Trust, UK)*
Dr. De Silva outlined a vision for the investigators in attendance on behalf of the Wellcome Trust: to create a globally connected network of large cohorts with multi-dimensional data from representative samples of diverse populations in hopes of maximizing the utility of our resources to improve health. She outlined steps for achieving this vision:

1. *Breadth of resources with representative samples from diverse populations*
   Much of the existing data is focused on white, northern hemisphere populations, which may not be useful for interventions that are impactful to the entire globe. There is a risk of misusing the resources as they grow larger (e.g. generalizing data that is not representative). Funding may be prioritized on new resources in populations for which there is no data, enrichment of existing resources to be more representative, and creation of better methods to produce representative data. Dr. De Silva noted that the group should aim to create an interconnected resource that inhibits large studies from crowding out some of the smaller cohorts of data that might be more representative to create a cohesive patchwork of data.
2. *High quality multi-dimensional data across life course*
   A patchwork of resources can achieve rich life-course data to include social, cultural, and environmental data in addition to biology and genes. Individual studies to collect this depth of longitudinal data are quite expensive.
3. *Data linkage, interoperability and analysis maximized across resources*
   It has been difficult to link routine data into longitudinal data sources, particularly with regard to ethical and consent frameworks. The data sets are not often interoperable, but the research community should strive to create data collection and data format standards. A collaborative sequencing strategy that perhaps prioritizes populations for which there is no data or chooses exome sequencing over whole genome sequencing would maximize funding resources. As these huge datasets are created, new methods for data analysis and linkage need to be evaluated, including machine learning, to look for patterns that may be generalizable.
4. *Resources are discoverable, accessible, and widely used*
   To the extent that it is legally possible, resources need to be discoverable and accessible. Individual cohorts and funders should encourage data sharing and progressive data access policies.
5. *Meaningful stakeholder engagement improves quality and impact of resources*
   There should be movement toward bi-directional dialogue and returning results to participants to foster public trust and develop rules of engagement with industry such that public trust is not jeopardized.
6. *Improved strategies for translation of findings into health impact*
   Cohorts are not all discovery science, and some of them have direct relevance for clinical practice. The pipeline of translation should be streamlined, perhaps through funding from national governments, creation of a coherent pipeline to policy and practice, and early engagement with policy makers.

Dr. De Silva explained that both funders and researchers have responsibilities to support this vision. The funders must provide the right opportunities to support the vision to ensure diversity and scale as well as drive regulatory and ethical frameworks for data sharing. Researchers must commit to this vision as well and use their own networks and influence to align methods and data collection strategies. Wellcome Trust has taken on a Longitudinal Population Studies Strategy to support these goals, drive resource standards, enhance the value of their resources, and work with other funders toward the common vision.

Innovations in Cohort Study Methods in All of Us – Kelly Gebo *(All of Us Research Program, NIH, USA)*
The All of Us Research Program plans to enroll over one million people, follow them longitudinally over 10 years and collect data from a variety of resources. They are using several innovative approaches to achieve their goals, which may set the stage for future large cohort studies. All of Us has engaged with a number of healthcare organizations and community partners to reach a diverse group of participants. Subjects provide responses to surveys, baseline physical measurements, biospecimens, and mHealth data from their personal Fitbits and potentially other wearables in the future.

All of Us has created a disease-agnostic scientific framework that should enable research to

- Increase wellness and resilience
- Reduce health disparities
- Improve risk assessment and prevention
- Provide earlier diagnosis
- Improve treatment and outcomes

Since the launch in May 2018, over 133,000 participants have completed the full protocol across all 50 states in the U.S. Initial demographics show a large amount of racial and ethnic diversity and some variation in gender identity. Over 75 percent of the enrolled population is considered underrepresented in biomedical research. In 2019 and 2020, the program will launch a Public Data Browser, a researcher workbench and begin genomics work. They are piloting a program to return results to their participants. They anticipate challenges with data sharing and matching to other publicly available datasets, selecting variables for disease-agnostic studies, allowing open data access without enabling bad actors, returning results to participants, and retaining long-term participation.

Overview of Current Progress – Geoff Ginsburg *(Duke University, G2MC, USA)*
Dr. Ginsburg summarized the historical context of IHCC and its goals. Individual cohorts may be limited by geography, size and ancestral origins, which constrains the questions each cohort can answer alone. In 2015, NIH compiled a list of large cohorts, and in June of 2017, the Heads of International Research Organizations (HIROs) agreed to bring these cohorts together to encourage data sharing, improve efficiencies, and maximize investments. In March 2018, the first cohorts summit was held at Duke University (USA), hosting 100 attendees from 24 countries that represented 60 cohorts that encompassed about 30 million participants. Cohorts invited to this summit had more than 100,000 participants (or plans to enroll this number), were not selected for disease, had biospecimens available, and had the potential for longitudinal follow-up. The summit solidified a vision for success: to create a global network for translational research that will utilize large cohorts to enhance understanding of the biological and genetic basis of disease and to improve clinical care and population health.

Since the initial summit, meeting materials have been published on the G2MC website, a summit white paper was drafted for publication, and three teams were formally established to focus on:

- Data standards and interoperability
- Scientific strategy and cohort enhancements
- Policy and bio-data sharing

Earlier in 2019, IHCC opened a Request for Ideas (RFI) for novel cross-cohort scientific programs from members. Many of the ideas from the RFI submissions are reflected in the summit agenda: rare conditions, exposures, and genotypes; consanguinity and founder populations studies; and other unique training opportunities. With 115 attendees from 25 countries and several new cohorts since 2018, this summit will emphasize a variety of collaborative opportunities and support the formation of data architecture, a common scientific agenda, and robust policy agenda.

Session 2 – Opportunities for Partnership
Chair: Peter Goodhand
10:30 AM – 11:15 AM

Cohorts in the EU and Beyond - Barbara Kerstiëns *(European Commission, Belgium)*
Dr. Kerstiëns gave an overview of some of the cohorts currently funded by the EU to include both disease-specific and longitudinal cohorts. To be funded by the EU, at least three institutions from three different EU member states must participate; on average, the projects include 9-11 partners. Applications are welcomed from all areas of the globe as long as they include three countries within the EU. In the last five years, there have been 233 collaborative projects of large and small cohorts. Collaborations have included work on breast cancer risk and screening (BRIDGES and B-CAST), Zika across 18 countries (ZIKAlliance), environmental interactions (Human Exposome Project), and an initiative to establish common infrastructure for national cohorts in Europe, Canada, and Africa (CINECA).

Collaboration between the cohorts presents certain challenges: understanding what resources are available and federating these, maintaining sustainable funding, providing and developing infrastructure and tools, and sharing data. SYNCHROS was funded to integrate stakeholders in the discussion of specific cohort issues. In 2021, EU will start a Horizon Europe framework program with the funding streams listed below. They will put out specific calls for proposals on these themes in 2020.

- Health throughout life course
- Environmental/social health determinants
- Non-communicable and rare diseases
- Infectious diseases
- Tools, technologies, and digital solutions for health and care
- Health care systems

IHCC will have the ability to influence the development of future funding streams. In 2018, the EU initiated a commitment to enable access via existing and future genomic databases to at least one million sequenced genomes in the EU by 2022. The requirements to achieve this goal mirror the vision of IHCC.

Moving from Genome Discoveries to Disease Mechanisms – Nancy Cox *(Vanderbilt University Medical Center, USA)*
Dr. Cox outlined an ongoing effort with the Common Disease Consortium (CDC) to understand the mechanism of common disease, moving from genomic discoveries to cellular mechanisms and on to disease mechanisms. Over the last 20 years, the discoveries have rarely led to an understanding of the driving mechanisms of disease. After building infrastructure, creating maps of genetic variation, and providing outputs of some foundational projects like GTex, there is now some insight into the mechanisms and pathways of disease. Pleiotropy – a mutation in one gene will affect the related cell type in all organ systems where it exists – is now recognized in cellular mechanisms. It is less often

recognized with regard to common disease mechanisms; this, however, would provide the opportunity to create larger targets for discovery. The consortium had an initial meeting in Oxford in December 2018 with 85 participants. They created working groups on human variation, systematic catalog of variants and function, disease-specific mechanism, data platforms, and data sharing and ethics. Thirty individuals are participating in an organizing committee to build this consortium with meetings scheduled in May and September 2019.

The European Project SYNCHROS – Josep Marie Haro *(Parc Sanitari Sant Joan de Déu, Spain)*
Dr. Haro echoed the charge given by President Jóhannesson to embody hope, modesty and duty. The SYNCHROS initiative was proposed in response to an EU call to build international efforts on population and patient cohorts. Given the rich variety of cohorts from Europe and other parts of the world, the level of integration needed to network these cohorts required escalation. SYNCHROS has set out to create a global, universal approach to optimize cohort study data by means of a strategy for integrating European and international health cohorts that are population-based patient and clinical trial cohorts aimed at prioritizing best practices across Europe and the rest of the world. SYNCHROS currently has 11 partners across six EU countries with external scientific and ethical advisory boards. The initiative's work packages include mapping the cohort landscape in Europe, identifying best strategies for integration, promoting data harmonization with standards, fostering the inclusion of data from new technologies, promoting best practices for access to cohorts, and contributing to an international strategic agenda for global cohort coordination. The project has three phases: 1) collecting and analyzing evidence, 2) coordinating methodology, and 3) creating and disseminating a sustainable strategy.

Session 3 – Data Standards and Infrastructure
Chairs: Phillip Awadalla & Thomas Keane
12:15 – 2:20 PM

Vision and Challenges to Achieve Cohort Interoperability – Philip Awadalla & Thomas Keane *(Canadian Partnership for Tomorrow Project, Canada; GA4GH, European Bioinformatics Institute, UK)*
Dr. Keane outlined the vision of IHCC's Data Standards and Infrastructure Team: to create interoperable IT standards and infrastructure, which will enable population-scale genomic and biomolecular data accessible across international borders. Genomic data carries many challenges; it may be 10-12 gigabytes for a single sample, and IHCC cohorts are mostly 100,000 participants or more. The metadata has heterogeneous phenotypes, various collection protocols and data dictionaries, which are then combined with different sharing restrictions across international lines. A federated model will provide benefits in which variables could be updated at a central hub and there may be flexibility for the cohorts to participate in other initiatives. The goals of GA4GH align well with the Data Standards and Infrastructure Team; they have well established data work streams with products rolled out in real world driver projects.

Dr. Awadalla gave an overview of the challenges associated with the interoperability vision of the Data Standards and Infrastructure Team. The most fundamental barrier is finding the data and creating a cohort atlas. They began by looking at Maelstrom, a catalog of mostly population cohorts, cross-referencing these with the IHCC cohorts, and collecting data dictionaries through a recent call to the IHCC cohorts. In an existing catalogue, the variables were collected, defined and categorized, and then evaluated for the harmonization potential.

The Genomic Aetiology of Osteoarthritis – Ele Zeggini *(Institute of Translational Genomics, Germany)*
Osteoarthritis is a common condition with no curative therapy, so there is a need to understand the mechanism of disease to develop disease-modifying treatments. Dr. Zeggini explained that the condition is caused by an interplay of genes, the environment and a number of epidemiologic risk factors; however, the pathogenesis of the disease is not fully understood. Dr. Zeggini discussed an ongoing project to improve understanding with a bigger sample size, better phenotype definition, and molecular profiling.

With the UK Biobank and UK National Joint Registry, they were able to achieve larger sample sizes for discovery and replication (50K, 260K, respectively). These studies established more associated loci and several genetic correlations between osteoarthritis and epidemiologically linked diseases like obesity and smoking. Next, they solicited biospecimens (saliva) from participants on the registry with hip dysplasia. The collection and discovery was low-cost and time-efficient. GWAS on these specimens revealed significant signal. The next step was to establish Genetics of Osteoarthritis with more cohort partners who together include 180,000 cases and 1 million controls. In time, they will continue to work to increase sample size and improve diversity in this collective cohort and evaluate the full allele frequency spectrum.

In defining a better phenotype, Dr. Zeggini acknowledged that osteoarthritis is an endpoint from various mechanisms and not one single disease. They have begun to examine a pathway to understand the mechanism of action, and from this, have initiated rapid-throughput phenotyping of knockout mice Osteoarthritis poses a unique challenge in molecular profiling, though, because the affected tissue is not easily accessible. They have collected knee tissues from those who have had replacements and examined the intact and degraded cartilage. This work has revealed candidate biomarkers for disease progression, disease stratification criteria and more.

Combining clinical information with -omics data has resulted in a better phenotype definition, novel genomic associations, molecular maps of the affected tissue, an understanding of disease progression, functional variants and mechanistic models; all of which have created an environment for translation.

RFI Presentations

Genetic and Non-Genetic Risk Factors for Uncommon Cardiometabolic Conditions - Adam Butterworth *(University of Cambridge, UK)*
The goal of this project idea is to study uncommon and rare conditions using the strength of multiple large cohorts within IHCC. This would enable the group to produce several rapid and impactful scientific manuscripts with existing data and demonstrate the utility of IHCC. Dr. Butterworth gave an example of how this is currently being done for hemorrhagic stroke with multiple cohorts, including some in IHCC. Although there were insufficient numbers of cases in an existing collection of 68 small-sized cardiovascular cohorts containing fewer than 100,000 participants, his group identified 10 larger cohorts with genetic data on subarachnoid hemorrhage and detected genome-wide signals that would not have been detectable in any one cohort alone. The proposed project would investigate another three uncommon conditions over the next two years with each involved cohort running a common analysis on their own data. The cohorts that would like to participate need only to have ICD code data and some genetic data, or ICD code data for looking at non-genetic risk factors. This project could be done at a low cost with centralized work from a few post-doc staff members with the analysis efforts of the participating cohorts. The next steps are to identify the phenotypes for exploration, ensure a shared

statistical plan can be applied across the cohorts, and determine a way to incentivize the cohorts to participate in this quick-win for IHCC.

A Pilot Study of Data Harmonization and Rare Variant Detection among International Cohorts – Rongling Li *(National Human Genome Research Institute, USA)*
Population health is affected by genetic and environmental factors. Having international cohorts gives the group an opportunity to study population health across different ancestries with varied health and environmental factors. Dr. Li introduced a project to examine how international cohort datasets could be combined for global health research. This pilot project would have four specific aims: First, standardize and harmonize cohort data sets to perform analysis, which would involve imputing existing genotype data from cohorts and perform whole genome sequencing of samples from underrepresented populations as needed. Second, collect high level environmental and location-specific data (e.g. weather, air/water quality, health access, etc.) from cohorts as needed. With this cross-cohort data set, aims three and four would compare frequencies and distribution of genetic variants across cohorts and their ancestral populations. Specifically, loss-of-function variants and clinically actionable variants would be evaluated. They would also investigate rare variants associated with selected clinical phenotypes with differences in prevalence across ancestral groups. The pilot study would utilize the work of the Data Standards and Infrastructure Data team and the Policy and Bio-Data Sharing team on data sharing and harmonization. Cohorts would analyze their own data and add it to a federated database with non-downloadable access for investigators. In year one, data harmonization and additional data collection would occur, followed by data cleaning and analysis in year two, and interpretation, publication, and dissemination in year three. This project would create an overarching framework for many of the studies IHCC could perform. The three-year pilot study would cost an estimated $18 million, with about one-third of that dedicated to sequencing of underrepresented populations.

High-Throughput Metabolomic Biomarker Measures across IHCC Cohorts - Hákon Hákonarson ***(Children's Hospital of Philadelphia, USA)***
Dr. Hákonarson explained that information from electronic health records and genomic, transcriptomic and epigenetic data sources could be combined with proteomic and metabolomic data to improve risk prediction and prevention of rare and common diseases. They have begun using a platform of 228 biomarkers (lipids, ketones, fatty acids, inflammatory markers, etc.) that at a large volume can be performed for less than $20 per sample. At the Center for Applied Genomics at the Children's Hospital of Philadelphia (CHOP), they are able to recruit randomly across multiple sites (population-based) for large biobanked pediatric cohorts. Pediatric subjects provide a unique opportunity to study the genetic influence of disease when environmental factors have had limited influence.

They would like to conduct a pilot study of 10,000 children with obesity to measure the biomarkers of this at-risk population and compare these with markers of adults in other cohorts with similar ancestry. This cohort would then be combined with other cohorts in IHCC with plasma and/or serum samples available to perform metabolic analysis. Combining metabolomics profiling with genomics using biobank cohorts can reveal early molecular insights into drug target biology, characterize pleiotropic effects and discover novel drug candidates through PheWAS approaches. The group would like to start with a pilot phase for the first year, measuring biomarkers across their existing cohort. In year two, they would integrate metabolomics with electronic health records, genome and risk factor data, and perform analysis. In year three, they would work on publishing preliminary data and expanding to phase two: integrating cohorts from Africa, South Asia, and South America. The project would cost an estimated $500,000 for Phase 1. Toward a Federated Data Ecosystem – Anthony Philippakis *(Broad Institute, USA)*

Dr. Philippakis outlined an inverted model for data sharing, where the data stays in one location and the researchers come to the data (cloud-based; it is not downloaded or copied), which supports data security and efficient use of resources. This would not be intended to create one monolithic entity to govern the world's data, and as such, it would need to be a federated ecosystem to combine diverse datasets. Dr. Philippakis suggested building this ecosystem in a modular way, so each group is still able to control their own data and who has access to it. This ecosystem would also need a workspace -- an analytical sandbox with a methods repository for analytical tools and collaboration. The federated model needs to be modular, open (source, API), community driven and standards based. They have already piloted some of this work between partners at the All of Us research program. This proposal would follow on that progress by combining the data from three cohorts: UK Biobank, Singapore 100K, and All of Us, to build a federated model that would work between the three of them. Each of these cohorts will have sequencing efforts underway in the next two years, and downloading these data will be untenable. Thus there is a great need to set up an effective model for sharing and cross-cohort analysis. Starting with these three cohorts is a great driver project over three years that would then set up for expansion of the model to other cohorts.

Session 4 – Scientific Strategy and Cohorts Enhancement
Chairs: Adam Butterworth, Hákon Hákonarson, & Gad Rennert
2:35 – 4:40 PM

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

Dr. Butterworth outlined the vision of the Scientific Strategy and Cohorts Enhancement Team to identify novel approaches to diagnose and treat common and rare conditions, explore enhancements to existing cohorts, and address diversity gaps. To achieve this vision, they aim to crystallize the emerging themes with a scientific strategy for IHCC, identify high priority questions that might be answered by IHCC, and refine pilot proposals to put forth to funders. The team solicited a request for ideas (RFI) from IHCC and received 23 proposals, many of which are being presented during the summit. Common themes from these proposals included harmonization, loss of function and human knockouts, polygenic risk scores, rare and uncommon conditions, precision medicine, combining existing data, proteomics and metabolomics, pharmacogenomics, and sequencing and genotyping. Challenges include lacking a framework for data sharing, prioritizing enhancements between so many cohorts, having uneven geographic and ethnic coverage, and determining how to harmonize the data. Cohorts should consider how they might like to participate in confronting these challenges and troubleshooting solutions.

FinnGen & Global Biobank Analysis – Mark Daly *(Institute for Molecular Medicine Finland, Finland)*
Dr. Daly described a unique ecosystem for genomic research in Finland, which inspired the FinnGen project. Finland has an efficient registry system, universal healthcare, an ethnically isolated population, enforced laws about medical record keeping, and rich resources on the population's genetic history. When an individual is recruited for a study, they also consent to have their sample linked to an entire lifetime of medical history to give longitudinal, rather than cross-sectional, data. FinnGen is a public-private partnership with two types of cohorts: existing cohorts with samples from 20-30 years ago that allow for follow up and prediction activities, and new clinic-based activities in specific disease areas.

Since August 2017, 50,000 subjects have enrolled with data freezes every six months. Preliminary data showed allele signals for inflammatory diseases, specifically for alleles that are common in Finland but rare in the rest of the world. They used gnomAD to examine the allele frequencies and found 15-20K protein coding variants testable in Finland's population but rare in the rest of the world. With this, they have made insights specific to Finland's population and examine common endpoints from these alleles, like type 2 diabetes. In the future, Dr. Daly's group would like to initiate a global biobank meta-analysis of about 2 million samples with only easily sharable summary statistics. This initial lightweight data sharing could take place this year on samples that are already analyzed, while IHCC and other groups pave the way for more substantial data and sample sharing. An initial endeavor on this with four cohorts identified 52,000 asthma cases and identified 52 independent loci; there are many other opportunities in the future with cohorts that would like to participate.

Body mass index and mortality: using big data for common conditions – Emanuele Di Angelantonio *(University of Cambridge, UK)*
There is a high prevalence of obesity in many countries globally, with 1.3 billion people overweight and 650 million obese. Dr. Di Angelantonio outlined an investigation linking being overweight or obese with excess mortality. Past research suggested that overweight status is not associated, but obesity is. There is a Global body mass index (BMI) mortality collaboration of 239 prospective studies with 10.6 million people from European, North American, and Southeast Asian regions. They tried to reduce bias by excluding pre-existing conditions, omitting initial follow-up years, and analyzing people who never smoked, as this is a known obesity confounder. Dr. Di Angelantonio's group found an association with BMI and the all-cause death rate ratio; the hazard ratios were higher on both ends of the BMI extremes (high and low). The associations were stronger in men versus women, younger adults versus older, and cancer cases versus respiratory or coronary heart disease cases. They validated this observational data with genetic data from the UK Biobank and confirmed that the shape of association was similar. Dr. Di Angelantonio concluded that analyzing individual participants from large scale cohorts can help address relevant public health questions related to common exposures and outcomes.

RFI Presentations

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)*
This project aims to look beyond host genetics and investigate the intersection between the host and the environment on circulating proteins and metabolites. Existing research has revealed that circulating proteins vary as a function of age, exposures, and health status, and many of them are associated with disease risk. Branched chain amino acids (BCAA) are associated with adiposity; furthermore, circulating BCAA have been shown to be associated with both diabetes risk and pancreatic cancer. This suggests that the circulating proteins may be of use in studying rare and common biology in the context of a larger consortium. Dr. Eliassen suggested they may be able to evaluate proteomic and metabolomics profiles and their impact on disease etiology across demographic groups, eventually leading to precision medicine approaches to prevention and treatment.

Using diabetes as a driver project, they aim to investigate the differences in proteomic and metabolomic profiles across populations, investigate how lifestyle factors influence these profiles, and examine their role in developing a chronic disease. They plan to select 2,000 disease-free controls and 1,000 incident cases of diabetes across several globally spaced cohorts matched as case-controls. They will perform proteomics on the OLink platform (proximity extension assay) and metabolomics on the Broad Institute's LC-MS semi targeted platform. Dr. Eliassen outlined a three-year project that would first define a manageable dataset between the cohorts, create a central data repository for harmonization, then select the cohort, assays, and analysis. They anticipate challenges with sample shipping, data harmonization, and creation of the central data portal. The project will require participation from cohorts across many geographies. Dr. Eliassen estimates a budget of about $2,000 per sample, plus funding for personnel at each cohort and central research staff.

Pharmacogenomic Analysis across IHCC Diversity Cohorts for Assessment of Drug Response - Kenny Nguyen and Hákon Hákonarson ***(Children's Hospital of Philadelphia, USA)***
Dr. Nguyen outlined several methods for utilizing large data sets to advance pharmacogenomics with deep learning. Deep learning and artificial intelligence methods can be used to extract knowledge and patterns in large data sets to predict drug responses and optimize drug selection and dosing. The model must be developed, trained and validated. Then large amounts of data can be analyzed using renormalization and coarse graining, which extracts relevant features of a physical system to describe phenomena at large-length scales by integrating short distance degrees of freedom. Limitations in computational power are directly related to the size of the molecular system. Solutions include modifying the software implementation or modifying the system representation to make the calculations more tractable. Quantitative systems pharmacogenomics could then be utilized to examine relationships between drug, biological system, and disease process. This would integrate quantitative drug data with knowledge of its mechanism of action, which would predict how drugs modify cell signaling networks and how they are impacted by human pathophysiology.

Dr. Nguyen proposed a three-year project to build, train and validate a deep learning model, which would then be expanded to large internal data sets, and eventually, larger cohort data. They aim to elucidate and implement novel pharmacogenomics associations and trends from drug response assessments using these methods. They will need to collaborate across cohorts in order to have enough data for discovery, reproducibility, and validation. The personnel and equipment requirements should cost about $750,000 USD each year for three years.

International 100k+ Consortium (IHCC) Drug Development Resource - Aroon Hingorani *(University College London, UK)*
Dr. Hingorani began by highlighting the unique opportunity for the consortium to focus resources on drug development and repurposing based on human genomic and linked biomedical data analyzed at scale. Using human genomics would help lower risk during drug development and reduce the high rate of drug development failure (96 percent) and subsequent increased drug cost to patients and governments. The high risk of failure encourages the pharmaceutical industry to take risk-averse approaches to avoid developing risky compounds and mechanisms of action. Genomics enables industry to work on the correct therapeutic targets and mechanisms of action and may also focus development pipelines on therapeutic needs that would benefit patients and societies globally. It is difficult to match the correct target with the disease, as they are most commonly studied early on with cellular and animal models, which may be unreliable predictors of treatment efficacy in humans. This often results in false discoveries that are taken forward to clinical phase testing. A genomic approach to drug development starts with study of the correct organism (humans) and allows for study of targets across the entire genome with low false discovery rates. The random allocation of alleles at conception emulates the design of a randomized control trial. Examining existing GWAS data showed rediscovery of licensed drug targets (approximately 80). Phenome-wide association studies have shown the benefit of target variation and repurposing; for example, the drug target IL6R for Rheumatoid Arthritis was shared by a wide range of other inflammatory conditions.

Over a 2-3 year timeline, this project would create an IHCC genomic data-driven drug target identification and validation pipeline, with a standardized approach to data generation, meta-analysis, visualization, reporting, and sharing in a form that is useful for drug development programs in academia and industry. Cross-cohort collaboration would be valuable for case identification and achieving proper scale, breadth and resolution of phenotypes. The resources required will depend on the scale of the initiative and whether additional -omics are needed.

The Human Knockout Project - Daniel MacArthur *(Harvard University/Broad Institute/Massachusetts General Hospital, USA)*
Human genetics provides a natural dose-response curve to evaluate target function and biological phenotype. Loss-of-function (LoF) alleles provide complete disruption of protein coding genes, which can be particularly informative as they provide an in vivo model of lifetime systemic inhibition of the target gene. Given known mutation rates and seven billion people, it is almost certain that every possible single base change compatible with life exists in a living human. In order to find human knockouts, we need to look beyond European populations to populations with different properties that increases the probability of a rare homozygous variant being present. Dr. MacArthur outlined two alternative strategies for knockout discovery. First, they could focus on bottlenecked populations, enriching for a set of variants that have passed through a population bottleneck and then become common in the population. This often leads to many samples per gene with which you can do robust statistical calculations on the association between these LoF variants and phenotypes. Second, they could focus on consanguineous populations (high level of parental relatedness), which increases the probability that a rare heterozygous variant is homozygous in the child. Their optimal strategy would be to integrate data across multiple populations, focusing on both bottlenecked (e.g. FinnGen) and consanguineous (e.g. East London Genes and Health, PROMIS) populations with hopes of using sequenced cohorts with access to deep phenotype data and genotype-based recontact.

Dr. MacArthur outlined the proposal goals using only existing exome and genome data. They would first produce a catalog of LoF variants from cohort sequencing data and develop the support frameworks required for large-scale global recontact experiments. The group would recontact sequenced individuals worldwide who are heterozygous or homozygous for LoF variants, consent and phenotype them, and eventually develop a public database of LoF variants. This Human Knockout Portal would build on experience with the gnomAD browser and display all known human LoFs from collected cohort data. The initial phase of the project would last three 3 years and require $200,000 per year for individual cohorts and central funding of $1.8M per year for data management, analysis, etc. Phase two would last five years and would include funding for sequencing selected cohorts and additional analysis, recall experiments, and the LoF portal.

Identification of Loss of Function (LOF) Variants in Health and Disease - Hákon Hákonarson ***(Children's** Hospital of Philadelphia, USA)*
Historically, LoF variants have guided novel developments for chronic disease and can guide biological direction of the respective pathways involved. Dr. Hákonarson highlighted a rare disease program at the Children's Hospital of Philadelphia with numerous LoF variants with GWAS. He used the example of DcR3 affecting the immune system and the irritable bowel disease (IBD) phenotype. When the Decoy Receptor is missing, LIGHT binds to its receptor instead, activating an immune response and creating inflammation, as seen in patients with IBD. Dr. Hákonarson outlined the opportunity with IHCC to replace LoF gene DcR3 in IBD with monoclonal antibodies. He outlined multiple proof of concept examples for rare LoF variants being imputed into untyped data. The pharmacogenomics model for LoF variants would start with the disease of interest, perform genetic screening, stratify disease with

biomarkers, perform intervention at biochemical pathways, and develop new drug candidates that would be fast tracked to clinical trials and market testing.

SESSION 5 – POLICY AND BIO-DATA SHARING
CHAIRS: GUN PEGGY KNUDSEN & LAURA LYMAN RODRIGUEZ
4:55 – 7:10 PM

Recap of Analyses to Date and Planning for Next Steps – Gun Peggy Knudsen & Laura Lyman Rodriguez *(Norwegian Institute of Public Health, Norway; National Human Genome Research Institute, USA)*
Dr. Rodriguez began by acknowledging an unequal representation on the Policy and Bio-Data Sharing Team from large cohorts in well-studied areas of the globe and invited IHCC members from smaller cohorts or other parts of the world to join them to ensure the policy agenda will be shaped with broader global needs in mind. The team's objective has been to develop a policy agenda to facilitate and promote assembling cohorts by identifying common needs, considering challenges to be addressed and producing options for a policy framework supportive of data sharing. Prior to the March 2018 summit, many cohorts provided policy-relevant information about their studies but with varying levels of detail and answers that sometimes appeared to contradict available published policies or known laws and regulations.

Therefore, in February 2019, the Policy and Bio-Data Sharing Team distributed a second survey to cohorts with more specific questions on data sharing. About a quarter of the 105 cohorts in IHCC responded, including several from underrepresented locations and a variety of cohort sizes. The second survey queried four different levels of data sharing: individual-level, summary-level, metadata, and specimens. Specimen sharing has the most stringent controls and oversight for sharing. About 90 percent of the cohorts would require data access committee (DAC) review for projects including individual level data or specimen sharing. This appears to be an independent review process from the IRB/Ethics Board review, which was also required for individual level and specimens in 63 percent and 73 percent of cohorts, respectively. Collaboration was also commonly reported to require co-authorship, but less so with metadata-only sharing. Fees for cost recovery with sharing were expected by a majority of cohorts for individual level and specimen sharing. Study consent documents allowed for data and specimen sharing for secondary studies in 22 of 24 and 21 of 24 cohorts, respectively. Dr. Rodriguez noted that while this information is helpful, it only represents 23 percent of IHCC's participating cohorts. She encouraged attendees to continue submitting policy survey responses, and/or providing their data sharing agreement documents or any website URLs for reference by the team.

The team plans to listen to the policy challenges raised within the RFI responses presented and identify common elements needed to develop a general data sharing framework to support the projects for the short- and long-term. They would also like to identify research questions related to policy effectiveness and discuss these further during the breakout sessions to develop goals for the next phase of work.

Large Scale Openly Accessible Blood-based Epidemiological Studies in India – Prabhat Jha *(University of Toronto, Canada)*
Dr. Jha began by charging IHCC to fill in the large gaps to understand human disease globally. His group is currently trying to do nation-wide mortality studies representing the population to identify key questions to pursue with prospective epidemiology. Using India as an example, there is large variation in overall and cause-specific adult mortality risks. Risk of death stratified by age (under 15 versus 15-69) has clear variation across districts. There were large differences between the North and South in the under 15 age group, 80 percent of which could be explained with known associations. However, in the

adult groups, the differences were starker between the East and West districts, only 30 percent of which could be explained by known associations (literacy level, smoking, alcohol use, etc.).

Dr. Jha's group wanted to understand these patterns and the unknown associations in India; they started by collaborating with the Indian Million Death Study (MDS) – a nationwide household mortality survey. This was important for India as many of the deaths occur at home and they wanted a true snapshot of mortality. The study included 1.4 million home visits and has thus far included 800,000 deaths.

Based on existing nationwide data, vascular diseases are the leading cause of death in middle age, with the dominant contributor as ischemic heart disease. Between 2000 and 2015, the risk of death from ischemic heart disease rose in both men and women, while stroke death risk fell. However, there were trend inconsistencies between districts and urban versus rural populations. There are some known risk factors for vascular disease such as smoking, blood pressure, and blood lipids, but factors like lacto-vegetarian diet and low body-mass index (BMI) are emerging.

Across India, low BMI is correlated with low literacy. While North America and Europe have a higher risk of mortality in low and high BMIs, in South India, the exposure risk is flat across low to high BMI. Dr. Jha's research has suggested this may be associated with lacto-vegetarianism, which is common in the Western/Northern districts. They found no reduction in overall mortality from this diet in either gender, and in fact, the wives in the surveyed households had an associated increase in cardiac risk. This was associated with men eating first in these households. Other risk factors include ambient air pollution; however, they have seen that applying Western exposure risks to the epidemiological factors in Indian populations may be skewing reported results. There are ongoing studies working to counter this concern (e.g. Indian Study of the Health of Adults in Barshi) that enroll entire households and collect biospecimens with long-term follow up. Moving forward, they hope to enroll approximately one million adults across 7-8 sites in hopes to fill the need for prospective blood-based epidemiology.

About European Research Infrastructures & Implementing the GDPR: Towards a Code of Conduct for Health Research - Michaela Mayrhofer *(BBMRI-ERIC, Austria)*
Dr. Mayrhofer first outlined research infrastructures, which are designed to facilitate and provide resources for research communities. There is a need to move away from the 3-5 year life cycle for projects in order to create a sustainable infrastructure for researchers. There are 13 resources in the European Strategy Forum on Research Infrastructures (ESFRI) roadmap, including Biobanking and Biomolecular Resources Infrastructure - European Research Infrastructure Consortium (BBMRI-ERIC). BBMRI is a federated research infrastructure facilitating the data and specimen collection for the benefit of human health. BBMRI has 21 member countries and several observers. They offer ethical/legal/societal guidance, IT tools, and quality management services.

Dr. Mayrhofer then discussed the General Data Protection Regulation (GDPR), intended to harmonize data protection rules in the EU. The rules reinforce data protection rights of individuals, facilitate the free flow of personal data in the single market and reduce administrative burden. Research is at the intersection of two needs in the GDPR: protecting the rights of individuals while enabling free flow of data. Research is still an exception to GDPR, but institutions still need to be prepared for GPDR. BBMRI began by creating a frequently asked questions document and formed working groups. Several EU countries are still preparing to implement GDPR despite a rollout in May 2018.

BBMRI hoped to provide guidance and clarity about GDPR with a code of conduct. Their code of conduct is intended to guide the proper application of the GDPR, clarify specific rules for data controllers in research, demonstrate compliance by controllers and processors and foster transparency in using personal data in health research. They brought together a drafting group from a variety of stakeholders and are working on texts but would like to collaborate with other groups drafting regulation in sectors other than research. The code will include lawfulness of processing, responsibilities for the controllers and processors, appropriate safeguards, anonymization of data, and practical examples. BBMRI is using European Data Protection Board (EDPB) provided guidance on creating codes of conduct. These codes will be useful as guidance for data protection across Europe but will not be a "holy grail."

RFI Presentations

Application of Polygenic Risk Scores (PRS) for Improved Health Outcomes in Alzheimer Disease, Cardiovascular Disease and Across multiple Neuropsychiatric-Phenotypes - Patrick Sleiman and Hákon Hákonarson (***Children's Hospital of Philadelphia, USA***)
Polygenic Risk Scores (PGS) are highly predictive of complex phenotypes with underlying polygenic inheritance. However, clinical implementation is limited, as much of the formative data is from European ancestry, and the scores are not portable across ancestries. Dr. Sleiman outlined this proposal's aims to delineate the heritability and generate PRS for various phenotypes for African American and other non-European populations and determine the extent to which different ancestries share PRS. This project would also explore the effects of age on prediction accuracy and examine the utility of PRS as biomarkers for targeted therapy.

Phenotypes of interest include cardiovascular disease, Alzheimer's disease, and neuropsychological conditions and developing them as biomarkers for drug discovery. PRS are developed with a large GWAS association analysis in order to develop multiple models to be validated in a large dataset for selection of the best fit model. Historically successful PRS have been associated with large sample sizes, most likely so they are feasible with European GWAS data sets, and therefore, only applicable to the same population. In examining age specific effects on PRS, heritability declines with age as environmental perturbations accumulate. Many of the GWAS samples that have guided PRS are from adults, and thus may underestimate risks in children.

Dr. Sleiman proposed collaborating with other cohorts in a three-year project to develop PRS with training and validation sets of non-European ancestry with age variation. This would require sharing samples across cohorts to gather adequate sample sizes with phenotypes of interest. The timeline will vary depending on the cohorts participating and the data they are able to provide. They plan to start by evaluating age effects in BMI, blood pressure and lung function, as this data is likely to be broadly available. They will create phenotype definitions and leverage electronic algorithms across the cohorts. Then they will perform analyses, which will lead to PRS being applied across the phenotypes. The project would need $300,000 - $400,000 funding for analytical personnel, effort for the participating cohorts to facilitate data transfer and collaboration, and data storage fees.

A Precision Medicine Approach to Multi-morbidity in Cardio-Metabolic Disease and Dementia - Sarah Bauermeister on behalf of John Gallacher *(University of Oxford, UK)*
Dr. Bauermeister presented this project idea using the Dementia Platform UK (DPUK) and including 47 cohorts that encompass over two million participants for data sharing. DPUK and IHCC share the aims of improving data access, identifying challenges associated with improving data access, and jointly finding solutions to these challenges. This project would use a precision medicine approach to multi-morbidity

as a use-case to identify and quantify governance and practical and scientific issues that help or hinder collaborative data access. The DPUK would be used as a trusted third party data portal that enables transparent collaborative research. DPUK allows for data sharing while enabling data controllers to maintain control of their data and restrict downloading. The platform categorizes the data types into a catalogue for quick reference of available data and allows members to apply for access to other cohort data. After standardizing much of the cohort data, DPUK has plans to harmonize a short list (30) of priority variables across all cohorts; these will be used for data visualization in an interactive tool. They hope this will assist participating researchers to better understand what type of resources are available in DPUK.

This two-year project would evaluate how an integrated analysis across data modalities and data sets could be achieved using standard regression and machine learning approaches. The group will produce a minimum viable product to test IHCC cohorts' ability to deliver multi-cohort, multi-modal data access. They will establish a consortium and find the areas of overlap, then standardize approaches for multi-modal data and establish work streams. The findings would be generalizable to IHCC to develop data standards and access policies.

Addressing the Ethnicity Gap in Human Genetics Research – Adam Butterworth *(University of Cambridge, UK)*
Dr. Butterworth proposed addressing the known ethnicity gap in human genetics research by conducting a large-scale (100K) pilot study to sequence substantial cohorts. Of existing GWAS data, 72 percent is from subjects recruited in the US, UK, or Iceland which leads to less utility of polygenic risk scores (PRS) across other ethnicities. Widespread genetic sequencing and phenotyping efforts in populations of different ancestries should reveal many more "gifts of nature." The existing cohorts are commonly of patients with clinical conditions in high income countries. There are some initiatives working to address this, but IHCC has a unique opportunity to include large, disease-agnostic cohorts. Well-powered rare variant association studies should involve discovery sets with at least 25,000 cases and a substantial replication set. Dr. Butterworth proposed a low-depth whole genome sequencing of 100,000 participant from LMICs, or a less expensive small-scale sequencing to develop population-specific genotyping arrays to apply at scale prior to imputation. He outlined the opportunity to study a Bangladeshi cohort, as South Asians are traditionally underrepresented but comprise a large portion of the international community. This LMIC does not have a national sequencing initiative, but it has a capacity building program (CAPABLE) devoted to this. There are options to study an imputed array (least expensive), whole exome sequencing, or whole genome sequencing (most expensive; most information). They hope to maximally benefit the studied communities with efforts, such as returning results to participants, maximizing the use of data, and facilitating the expansion of this effort beyond a pilot. The resources required are variable depending on performance of whole genome sequencing ($30-50M) or genome wide genotyping and ethnicity specific imputation ($3-5M).

Atopic Diseases, Early Life Exposure Risk Factors, and Genetics in the Young (ALERGY) Study – Elizabeth Jensen *(Wake Forest School of Medicine, USA)*
Dr. Jensen proposed a gene environment interaction study in pediatric cohorts, specifically in development of allergic disease. Allergic disease has increased in incidence and prevalence in many countries; one area of marked increase has been specifically eosinophilic esophagitis (EoE). Allergic diseases arise in the setting of underlying genetic susceptibility but are also the result of environmental exposures. While the symptoms are often managed by individual organ specialists, they often share a common pathomechanism (e.g. IgE mediated). Prior research has shown that early life presents a unique period of susceptibility to exposures and that early life factors are associated with the

development of allergic disease. Some mechanisms have not been studied for allergic disease, such as pro-inflammatory (e.g. maternal adiposity) or epigenetic mechanisms. Gene-environment interactions are valuable for identification of novel susceptibility genes and pathways for pharmacologic targets in disease treatment. The interactions studied should improve understanding of disease through recognition of disease heterogeneity; allowing for personalized treatment. These interactions may also identify novel, potentially modifiable risk factors among genetically susceptible individuals.

Gene-environment interaction studies have been challenging due to sample size constraints. Dr. Jensen outlined Phase I: the first aim would be to estimate the prevalence of allergic diseases and characterize the distribution of risk factors across pregnancy and early childhood cohorts. The second aim would be to estimate the association between risk factors, including demographic, health-related, geographic, perinatal, and early life factors, with childhood allergic diseases. They identified several cohorts to start with across the U.S., Norway, Denmark, Korea, and Japan, which would encompass more than 300,000 children, some with genotyping already in process. In Phase II, they would aim to evaluate gene-environment interaction in developing allergic diseases. They anticipate some data access and harmonization challenges and achieving adequate power for rare outcomes. Phase I would last for three years and will cost an estimated $585,000 in direct costs.

Day 2: April 24, 2019, 8:00 AM – 4:00 PM

Big Data in Health Care and Research, Opportunities and Challenges **–** Ewan Birney *(EMBL-European Bioinformatics Institute, UK)*

Dr. Birney gave an overview of the recent history of global genomics. In 2003, the sequencing of one genome cost as much as the most expensive home in London; now in 2019, it is affordable enough that many can have their genomes sequenced. As such, in the coming years, hundreds of millions of genomes will be sequenced. This increase in valuable genomic data has resulted in a shift in biological research out of the wet lab toward computer science. Similarly, genomic resources are shifting towards implementation more commonly in healthcare versus in research. On the healthcare side, there are deliverables expected for which practitioners will be held accountable; this is not the case in research. Moreover, while the subject of the research - human biology - is universal, human healthcare is quite diverse. This suggests that genomic research will not be implemented in a one-size fits all approach, and the path from research to medicine will vary. Healthcare is expensive but well resourced. Once a procedure has demonstrated its utility, it is usually rolled out for application to everyone; leading to a large amount of genomic data in the coming years.

To prepare for this, the Global Alliance for Genomics and Health (GA4GH) aims to become the driver body of standards for genomic medicine. Their mission is to accelerate progress in human health by helping to establish a common framework of harmonized approaches to enable effective and responsible sharing of genomic and clinical data, and by catalyzing data sharing projects that drive and demonstrate the value of data sharing.

GA4GH recognizes that the healthcare system is not used to the challenges associated with genomic data in terms of scale. Beyond one healthcare system, this data needs to be connected internationally in order to understand global diseases, particularly rare diseases. In order to work through these challenges, genomic data needs to be federated, have broad reciprocal access, have skill transfer to the healthcare setting, and include more discoveries. GA4GH has several driver projects and cohorts from geographically diverse locations. They have begun to develop technical standards to include the CRAM file format, Beacon API, Variant versus Binary Call Format, and several others. They are also working with other standards organizations like HL7 and ISO on shared goals. Dr. Birney concluded his talk by explaining that large-scale cohorts should be the drivers of these standards and invited the meeting attendees to become involved with GA4GH regardless of location.

Data Challenges in Sharing Across Cohorts – Nicola Mulder *(University of Cape Town, South Africa)*
**Note: This talk was originally scheduled in Session 3 on Data Standards and Infrastructure*
Dr. Mulder spoke about some of the data and IT challenges associated with cross-cohort collaboration using the example of H3Africa. She outlined challenges with data from the clinical and the research spaces, and how they specifically apply to the cohorts in H3Africa. For example, some of the medical records are not electronic, or uniformly kept. Data sharing has the existing challenges of differences in consent, data access agreements, interoperability, etc. However, data sharing and transferring with encryption is particularly challenging in Africa. This is coupled with a strong feeling of exploitation from African scientists from a historical past of helicopter science and challenges that still exist today on having resources that allow for rapid data analysis. Another challenge is having different ethical guidelines across 34 countries involved in H3Africa. Genetic studies have not commonly been conducted in all of these countries, so there are not clear laws and guidelines to start with. IT challenges include some lack of data-related skills that would enable rapid analysis and lack of bandwidth and storage resources. In the past, they have found it easier to ship hard drives from one site to another rather than attempt large downloads.

The H3Africa Initiative has focused its attention on alleviating these challenges for future studies by creating core phenotypes and a CRF template in REDCap that can be adapted to fit the local context. They have been working to map consent forms into data use ontology and create a searchable catalogue of cohorts that meets particular criteria. In the future, they would like to create and define more appropriate variable labels. For example the ethnic label of African does not allow for precision medicine approaches across 54 countries and ethnicities; however narrow labels such as tribal affiliations may reveal identifiable information despite anonymized data. Despite many of these hurdles, with time, the initiative hopes to make the data in H3Africa findable, accessible, interoperable and reusable.

Charge to Breakout Groups – Geoff Ginsburg *(Duke University, G2MC, USA)*
Dr. Ginsburg outlined the progress so far in the meeting and the mission for the breakout groups over the next several hours. There have been 14 ideas presented to the group for cross-cohort methods and collaboration, but there are several other great ideas in the meeting booklet that should be considered during the breakout sessions. The Data Standards and Infrastructure Team will focus on data standards and infrastructure, and continue their focus on data architecture and interoperability. In the future they hope to create a query-able data atlas for shared use for IHCC, and build on several of the similar efforts described during this meeting. The Scientific Strategy and Cohorts Enhancement Team will work on distilling a scientific agenda for IHCC from the many great ideas presented. They will aim to mold these ideas into an agenda that balances the scientific diversity of IHCC, but is also cohesive for funding consideration. The Policy and Bio-Data Sharing Team will discuss how to enable the activities and scientific agenda of IHCC with policy. Many of the proposals presented articulated significant policy challenges on data sharing, specimen sharing, regulatory landscapes, etc. that will be considered by this group.

Each team will guide discussion among the attendees of the breakout sessions and report back to the larger meeting group in the afternoon. The output of this work should enable our next steps in obtaining funding. IHCC leadership is planning to discuss our progress and plans with the Heads of International Research Organizations (HIROs) at their next meeting on June 17, 2019. As HIROs initially supported the formation of IHCC, the hope is that they may continue to support the group with concrete vision and goals. In the future, IHCC will also consider having interested industry partners as collaborators; introducing an alternative funding stream.

Session 6 – Team Break Out Sessions
8:45 – 10:45 AM

Each team listed below had group discussion that was summarized and reported to the larger group during Session 8. See the notes in Session 8 for details of these sessions.

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

Session 7 – Team Leads Synthesis
Breakout Session with Team Leads and IHCC Leadership
11:15 AM – 12:45 PM

After the breakout sessions, each of the team leads convened with IHCC leadership to report their progress from the breakout sessions and synthesize a path forward for the larger group. The outputs of this smaller group discussion was presented to the group in Sessions 8 and 9.

Session 8 – Break Out Reports and Discussion
Moderator: Teri Manolio
12:45 – 2:45 PM

Dr. Manolio started this session by reviewing the history of IHCC and the progress that led to the current meeting. At the first summit in March 2018, the organizers invited the cohort leaders they knew of, and at that meeting, identified several more cohorts that should be involved in the IHCC. After the initial summit, IHCC leadership met with HIROs who provided seed funding to stand up the work teams and host the next meeting in November 2018. The group decided to open a Request for Ideas (RFI) to inspire discussion of what IHCC could do as a group and guide the scientific agenda. This RFI is intended to be an iterative process moving forward. This first request was issued in January 2019 with a February due date and March review date. The favored proposals demonstrated innovative uses of cohort resources, had a broad scope, enabled quick-wins for IHCC, will potentially attract funders, were well-resourced with appropriate expertise, and possibly provide opportunities to young researchers. The short RFI proposals included a summary, structured abstract and a limited amount of additional information.

Proposal reviewers included IHCC team leads and the Executive Committee, who conducted an NIH-style peer review. The review process was explained by Dr. John Connolly from the Children's Hospital of Philadelphia, who assisted in the RFI process. Each of the 23 proposals received was assigned one primary and two secondary reviewers; review panel members were restricted from reviewing proposals they were associated with. The 13 reviewers scored the proposals across four categories: significance, approach, innovation and impact. The reviewers calculated the mean scores and ranked the proposals. The review panel convened to allow for discussion and changing of scores as needed. Although only some of the proposals were highlighted for presentation, each of them were included in the meeting booklet to catalyze discussion in the breakout sessions in hopes of piecing several ideas together for some flagship projects for IHCC.

Before starting the individual team summaries, Mr. Robert Eiss from the NIH Fogarty International Center described the structure and function of the HIROs. HIROs is an informal group without a website or terms of reference. It includes 29 of the major medical research councils and philanthropies and represents 90 percent of the global medical research spending. While the collection of funders in HIROs is not a funding or governing agency, they have been able to coordinate the movement of global projects. After recognizing a public market failure in non-communicable disease research in low- and middle-income countries (LMICs), the HIROs developed research questions collaboratively and began to coordinate budgets and supported scientists evaluating these research questions. They also underwrote the development of a descriptive database to enable collaboration between their institutions for efficient planning and implementation of studies in Africa. Dr. Francis Collins of NIH approached the HIROs to ask for a list of the large-scale (100K+) population-based cohorts they are funding. With a starting list of more than 50 cohorts, this supported the notion that a forum for these cohorts might be beneficial, leading to the inception of IHCC.

Data Standards and Infrastructure Team - Philip Awadalla and Thomas Keane *(Canadian Partnership for Tomorrow Project, Canada; Global Alliance for Genomics & Health; European Bioinformatics Institute, UK)*

Dr. Keane began the Data Standards and Infrastructure Team breakout report by presenting the team's revised vision: To deliver fair IT standards, best practices, and infrastructure to enable population scale genetic, phenotypic, environmental, and biomolecular data accessible across international borders accelerating research and improving the health of individual residents across continents. This vision incorporates *fair* IT standards – indicating findable, accessible, interoperable and reusable. The term "data" was also broadened to include types beyond genetic data. Dr. Keane then outlined the goals of the Data Standards and Infrastructure Team by defining a use-case of the end-to-end research cycle that will inform their work.

- Discovery: The cycle begins with data discovery and exploration of the semantic variables available for data harmonization. As the Data Standards and Infrastructure Team continues to collect data dictionaries from across IHCC, they hope to identify a minimum set of variables for harmonization.
- Access: Research questions should be defined with considerations of how to combine data across the involved cohorts (e.g. data use ontology). Research IDs will interplay with institutional and individual authentication and authorization through a data access controller.
- Federation: The federated data should enable research and/or clinical applications with analysis performed in the cloud. Data object services, tool registries and workflow execution environments will need to be brought to the cloud.

- Analysis Combination: When the analysis leaves the cloud, the results will be combined for application.

Dr. Keane outlined an in-depth discussion of researcher IDs and their importance for cohort interoperability. The researcher ID would be an assertion of a researcher's identify with respect to some sort of associated legal entity (e.g. university, hospital). The data access committee would need to set minimum requirements based on regulatory and legal requirements. The access structure should also define the interaction of citizen scientists without an associated entity. There are several research ID systems (e.g. Orcid, ERA, ELIXIR) available. The Data Standards and Infrastructure Team would like to work with a pilot project to establish authentication-authorization interoperability, perhaps with varied geographical engagement.

During the breakout session, the Data Standards and Infrastructure Team agreed on several proposals to develop further. Starting broadly, the Cohort Discovery Atlas would encompass as many cohorts as possible from IHCC to provide information on the most relevant cohorts with relevant phenotypes of study; making them discoverable. This is one of the most fundamental challenges for cross-cohort collaboration. Dr. Awadalla went into further detail about the Atlas, explaining that the team has been committed to this project since the initial IHCC Summit at Duke in 2018. They plan to characterize the cohort metadata to include phenotypes, genotypes, administrative health data, exposures, etc. and build on meta-domains in existing data catalogues (e.g. Maelstrom) and cross-harmonization projects (e.g. Tomorrow CHPT). With time, they would like to include more of IHCC cohorts with many more variables. After hearing some of the concerns from H3Africa and LMIC cohorts, the team will also evaluate ways to make participating in the data atlas valuable to those cohorts without taxing resources.

The second proposal would narrow the focus on a few narrow driver projects and take them through the research cycle from discovery, access, federated application to analysis. This proposal was presented on Day 1 of the meeting with the title: Towards a Federated Data Ecosystem. The team is hoping that this project will be a proof-of-concept and allow the same methods to extend concentrically to a larger number of cohorts. The team outlined a plan for this project over three years starting with one interface for the federated resources and a consistent system of access with analysis across at least three cohorts (e.g. UK Biobank, Singapore 100K, All of Us, European Genome-phenome Archive/H3Africa) in different data centers and/or regions. There was some discussion of how and when to most effectively involve LMIC cohorts in this project. While early engagement is targeted at the very large cohorts, waiting three years before working with smaller LMIC cohorts may result in some missed opportunity.

A third priority that the team did not fully explore due to time constraints is the development of a set of data best practices for IHCC. This activity would build on existing standards from groups like GA4GH while complimenting the activities of the Policy and Bio-Data Sharing Team. Guidelines might include practices for data security, recommended APIs for access and authentication, etc. The best practices would then be available for funders to mandate for future projects.

Scientific Strategy and Cohorts Enhancement Team - Adam Butterworth, Hákon Hákonarson, & Gad Rennert ***(University of Cambridge, UK; Children's Hospital of Philadelphia, USA; Carmel Medical*** *Center/Technion–Israel Institute of Technology, Israel)*

Dr. Butterworth outlined the progress of the Scientific Strategy and Cohorts Enhancement Team on crafting an IHCC scientific agenda using RFI proposals as points of discussion. The team started their breakout session by grouping the proposals into categories: using existing data and generating new data and/or enhancements. Projects generating new data may need to consider industry engagement to assist with genome sequencing, proteomics, metabolomics, etc. They heard from the pharmaceutical representatives in the breakout session that there should be industry stakeholder engagement in research design and development, prior to the proposal for funding. These projects may also consider the harmonization of clinical phenotypes, which was a high priority for pharma engagement.

The scientific agenda should be attractive to potential funders; IHCC cohorts are good resources as they contain population-based longitudinal data, enabling the study of many diseases and conditions. Initial proposed projects included recurring themes like polygenic risk scores, loss-of-function knockouts, and uncommon diseases. Initial IHCC projects should prioritize a wide inclusivity of cohorts; perhaps without requirements for genomic or molecular data, or by highlighting cohorts with epidemiologic data. Projects should demonstrate the cohesiveness of IHCC, perhaps by using ICD code phenotypes that can be applied across many cohorts.

The team plans to prepare a Scientific Agenda Summary Document. This will contain a pitch of what IHCC could achieve and outline its key selling points with the data and resources available. They will outline several key themes reflected in a short list of pilot projects. This document will be circulated to the group for their input before it is presented to HIROs.

During group discussion, Dr. Ginsburg noted that for this Summit, industry representatives were only invited based on their scientific contributions to IHCC cohorts. However in the future, IHCC plans to have more active engagement with industry as potential collaborators who can advance the science of the cohorts and benefit from shared interest in the research questions.

The group also considered research questions with epidemiological focus that could be incorporated into the scientific agenda. Multi-morbidity (multiple chronic conditions co-occurring in the same individual) in disease clusters across different populations would not require genetic data. The investigation of common exposures and risk factors across populations with a high level of sensitivity would be possible with very large data sets. Other suggestions included the timing of exposures, early prediction of disease from following healthy populations over time, social determinants of health, imputation of exposures, Mendelian randomization, etc.

Policy and Bio-Data Sharing Team - Gun Peggy Knudsen & Laura Lyman Rodriguez *(Norwegian Institute of Public Health, Norway; National Human Genome Research Institute, USA)*
Dr. Rodriguez acknowledged the varying perspectives in the Policy and Bio-Data Sharing Team breakout and explained that their discussion was framed around specific needs that would result from several of the proposed ideas. The team agreed on four directions of focus.

First, they would like to shift their focus from purely *data sharing* as the main goal to encouraging true *collaboration among cohorts* with data sharing as one of several outcomes of that goal. To articulate what can be gained through real collaboration, the team proposed to create a “first principles” document that would define at a high level the value of IHCC participation for cohorts. The principles would outline incentives, return on investment, demonstrated value to governing and funding bodies and individual cohorts, etc. It may outline options for capacity building, true collaboration with local scientists, an explanation of cost-sharing for work performed, among other concerns for participating cohorts.

Second, the Policy and Bio-Data Sharing Team recommended that IHCC presently prioritize collaborative proposals that only require federated analytic strategies. This could catalyze the development of a data sharing model and access framework that can involve more complex data sharing scenarios with time. They would like to start the discussion with the scientific teams to build around their research goals, and then evaluate the capacity of the involved cohorts to enable that scientific design based upon cohort input directly or through their submitted responses to the policy survey.

The third direction was to pursue policy specific topics that support the prioritized proposals. These topics may include return of results, particularly when this may be performed in programs across several nations with varying laws and regulatory frameworks. This direction might also include work to develop or identify best practices for industry engagement, publication policies, etc. They plan to iterate with the Data Standards and Infrastructure Team on how to include policy relevant information (legal structures, sponsorship, etc.) in the data standards and infrastructure.

Last, the team would like to collect data and information useful to inform policy development that IHCC and individual cohorts might pursue. They would like to gather knowledge from the experiences of each cohort participating in collective projects and their respective data sharing models to seize opportunities to collect real-time data of effectiveness or challenges of those models. As specific research proposals come forward, there may be policy research questions that could be asked in context.

Dr. Rodriguez concluded her summary by inviting IHCC members to join the Policy and Bio-Data Sharing Team and provide their input on policies that will affect the larger IHCC collective. She also requested that cohorts' members complete the policy survey re-circulated by Teji Rakhra-Burris during the Summit on behalf of the team to collect cohort-specific policy and consent details.

After a suggestion from the audience, the team agreed to provide some examples of consent form language that has been used in prior projects with planned international collaboration and/or sharing. There are some existing resources available that can also be shared with the group. Collaboration with international researchers will not always mean sharing the data and thus it may not always need to be specified in the consent.

With time, the team's work may contribute to an overall governance structure for IHCC. As this structure for IHCC will be developed democratically, the team will not need to draft this as a group, but could provide valuable comments and input on this deliverable. As work proceeds, it may also be valuable to have co-leaders from different time zones and perhaps multiple meetings to accommodate all participants.

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, & Teri Manolio *(Duke University; Global Alliance for Genomics & Health; National Human Genome Research Institute; USA and Canada)*

Dr. Manolio concluded the Summit by giving an overview of meeting objectives and progress towards these goals.

- Meeting Objective 1: Identify scientifically meritorious cross-cohort research projects and identify international collaborators willing to organize and participate in them.
  - A number of meritorious cross-cohort scientific projects were identified and presented.

  - International collaborators were not identified for all projects but some were engaged.
- Meeting Objective 2: Develop an IHCC scientific agenda to bring forward to funding bodies.
  - There was progress on this and these discussions will be formulated into a cohesive document and circulated for feedback.

Beyond the original objectives, this meeting revealed considerable interest in becoming a more formal consortium with charter, guiding principles, membership guidelines, etc. In the future, IHCC will consider a second RFI on a more relaxed timeframe that focuses specifically on projects that can only be done by an IHCC-like entity across many cohorts or on project with enhancements that benefit the entire cohort body. Several non-genetic cross-cohort research ideas were suggested during the Summit such as multi-morbidity, exposome/rare exposures, evaluating timing of exposures, social determinants of health, climate change, imputation of exposures from genomic/other data, and harmonization of non-genetic data.

As a group, the next steps will be to:

- Develop and distribute a summary of the meeting
- Develop exemplar proposals for presentation to HIROs from each IHCC team (Data Standards, Scientific Strategy, and Policy/Bio-Data Sharing)
- Bring priorities to the HIROs
- Develop a charter and guiding principles for the consortium, potentially including a more formalized sign up process and consider the expectation that Team co-leads cover the breadth of time zones to promote inclusivity
- Conduct a second solicitation of RFIs with sufficient advance notice, time to respond, and review of proposals by a broader section of the consortium
- Promote opportunities for trainees and junior investigators to participate in IHCC activities
- Circulate the list of cohorts in attendance and those that IHCC has knowledge of to ensure that no cohorts have been overlooked
- Based on the teams' and proposed projects' progress, host a third summit in 2020 with as many cohorts as possible represented as well as industry participation. This summit should allow presentations of progress from the ongoing projects.

While IHCC aims to find funding support for cross-cohort projects, traditional funding streams should still be considered for larger cross-cohort proposals. The standard funding mechanisms may be easier to achieve in the short term than a new programmatic effort. The Wellcome Trust has a collaborative award for up to $4 million for up to five years. Several of the projects presented at the Summit would qualify for these awards with the stipulation that they must be led by an institution in an LMIC or in the UK. Genome Canada, among others, has periodic calls for collaborative projects.

Attendees

Name	Affiliation	Country
Malak Abedalthagafi	King Abdulaziz City for Science and Technology	Saudi Arabia
Nahla Afifi	Qatar Biobank - Qatar Foundation	Qatar
Jesus Alegre-Díaz	National Autonomous University of Mexico	Mexico
Jessica Alföldi	Broad Institute	USA
Garnet Anderson	Fred Hutchinson Cancer Research Center	USA
Philip Awadalla	Canadian Partnership for Tomorrow Project	Canada
Sarah Bauermeister	European Prevention of Alzheimer’s Dementia (EPAD)	UK
Ewan Birney	The European Bioinformatics Institute	UK
Dan Brake	Sequence Bioinformatics	Canada
Adam Butterworth	University of Cambridge	UK
Lon Cardon	BioMarin Pharmaceutical Inc.	USA
Juan P. Casas	VA Boston Healthcare System	USA
John Chambers	Nanyang Technological University	Singapore
Zhengming Chen	Oxford University	UK
Tammy Clifford	Canadian Institutes of Health Research	Canada
Rory Collins	Oxford University	UK
John Connolly	Children's Hospital of Philadelphia	USA
Nancy Cox	Vanderbilt University	USA
Mark Daly	Institute for Molecular Medicine Finland	Finland
Mary De Silva	Wellcome Trust	UK
Joshua Denny	Vanderbilt University	USA
Emanuele Di Angelantonio	University of Cambridge	UK
Rajesh Dikshit	Tata Memorial Hospital	India
Sylvain Durrleman	INSERM/Aviesan Institute of Public Health	France
Mark Effingham	UK Biobank	UK
Margaret Ehm	GSK	USA
Robert Eiss	John E. Fogarty International Center	USA
Heather Eliassen	Harvard T.H. Chan School of Public Health, Harvard University	USA
Paul Elliott	Imperial College London	UK
Jonathan Emberson	University of Oxford	UK
Arash Etemadi	National Cancer Institute	USA
Catterina Ferreccio	Pontificia Universidad Católica de Chile	Chile
Tom Fowler	Genomics England	UK
Neal Freedman	National Cancer Institute	USA
Nobuo Fuse	Tohoku University Tohoku Medical Megabank Organization	Japan
Kris Ganjam	Microsoft	USA
Ya Gao	BGI	China

NAME	AFFILIATION	COUNTRY
J. Michael Gaziano	VA Boston Healthcare System	USA
Kelly Gebo	National Institutes of Health	USA
Christian Gieger	Helmholtz Zentrum München	Germany
Geoffrey Ginsburg	Duke University	USA
Roger Glass	John E. Fogarty International Center	USA
Marcel Goldberg	INSERM UMS 11	France
Peter Goodhand	Global Alliance for Genomics & Health	Canada
Eric Green	National Human Genome Research Institute	USA
Fran Grodstein	Brigham and Women's Hospital	USA
Joseph Grzymski	Renown Health/Desert Research Institute	USA
Hákon Hákonarson	Children's Hospital of Philadelphia	USA
Josep Maria Haro	Parc Sanitari Sant Joan de Déu	Spain
Aroon Hingorani	University College London	UK
Atsushi Hozawa	Tohoku University Tohoku Medical Megabank Organization	Japan
Birgir Jakobsson	Ministry of Health	Iceland
Rahman Jamal	Universiti Kebangsaan Malaysia	Malaysia
Sun Ha Jee	Yonsei University	South Korea
Yon Ho Jee	Yonsei University	South Korea
Elizabeth Jensen	Wake Forest School of Medicine	USA
Jae-Pil Jeon	Korea National Institute of Health	South Korea
Prabhat Jha	University of Toronto	Canada
Farin Kamangar	Morgan State University	USA
Norihiro Kato	National Center for Global Health and Medicine	Japan
Thomas Keane	European Bioinformatics Institute	UK
Barbara Kerstiëns	European Commission, Directorate-General for Research and Innovation	Belgium
Sung Soo Kim	Korea National Institute of Health	South Korea
Gun Peggy Knudsen	Norwegian Institute of Public Health	Norway
David Ledbetter	Geisinger Health	USA
Sarah Lewington	University of Oxford	UK
Rongling Li	National Human Genome Research Institute	USA
Chi-Ming Liang	Academia Sinica	Taiwan
Rachel Liao	Broad Institute	USA
Paulo Lotufo	University of São Paulo	Brazil
Beatrice Lucaroni	European Commission, Directorate-General for Research and Innovation	Belgium
Ryan Lui	BGI Europe	UK
Chris Lunt	National Institutes of Health	USA
Laura Lyman Rodriguez	National Human Genome Research Institute	USA
Daniel MacArthur	Harvard University/Broad Institute/Massachusetts General Hospital	USA
Per Magnus	Norwegian Institute of Public Health	Norway
Reza Malekzadeh	Tehran University of Medical Sciences	Iran

NAME	AFFILIATION	COUNTRY
Teri Manolio	National Human Genome Research Institute	USA
Tohru Masui	Keio University	Japan
Prashant Mathur	Indian Council of Medical Research	India
Michaela Th. Mayrhofer	BBMRI-ERIC	Austria
Hamdi Mbarek	Qatar Genome	Qatar
Martin McNamara	Sax Institute	Australia
Joe McNamara	Medical Research Council	UK
Mads Melbye	Statens Serum Institut	Denmark
Usha Menon	University College London	UK
Andres Metspalu	University of Tartu	Estonia
Takayuki Morisaki	University of Tokyo	Japan
Nicola Mulder	University of Cape Town	South Africa
Yoshinori Murakami	BioBank Japan/ University of Tokyo, Institute of Medical Science	Japan
Kenny Nguyen	Children's Hospital of Philadelphia	USA
Thea Norman	Bill & Melinda Gates Foundation	USA
Donna Parker	Duke University	USA
Alexandre Pereira	University of São Paulo	Brazil
Mauro Petrillo	European Commission	Belgium
Anthony Philippakis	Broad Institute	USA
Brittany Ploss	Duke University/Global Genomic Medicine Collaborative	USA
Erica Pufall	Wellcome Trust	UK
Teji Rakhra-Burris	Global Genomic Medicine Collaborative	USA
Gad Rennert	Carmel Medical Center/Technion–Israel Institute of Technology	Israel
Gabriela Repetto	Clínica Alemana Universidad del Desarrollo	Chile
Jessica Reusch	National Human Genome Research Institute	USA
Norie Sawada	National Cancer Center Japan	Japan
Catherine Schaefer	Kaiser Permanente	USA
Alan Shuldiner	Regeneron Genetics Center	USA
Patrick Sleiman	Children's Hospital of Philadelphia	USA
Anthony Swerdlow	Institute of Cancer Research	UK
Patrick Tan	Agency for Science, Technology and Research	Singapore
Guðni Thorlacius Jóhannesson	President of Iceland	Iceland
David van Heel	Queen Mary University of London	UK
Mara Vitolins	Wake Forest University	USA
Andreas Weser	Norwegian University of Science and Technology	Norway
Daisuke Yasumizu	Japan Agency for Medical Research and Development	Japan
Tsungfu Yu	Academia Sinica	Taiwan
Eleftheria Zeggini	Institute of Translational Genomics	Germany
Wei Zheng	Vanderbilt University	USA
Marie Zins	INSERM UMS 11	France

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 Nahla Afifi
- 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
- Nurses' Health Study
- Nurses' Health Study II
- Ontario Health Study
- Prospective Epidemiological Research Studies in IrAN (PERSIAN) Cohort
- Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO)
- Qatar Biobank
- Qatar Genome Project
- Saudi Human Genome Project
- SG100K
- Shanghai Men and Women's Health Studies
- German National Cohort (NAKO)
- Golestan Cohort Study
- 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
- Nord-Trondelag Health Study (HUNT)
- Norweigan Mother and Child Cohort Study (MoBa)
- 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

SPONSORS

Thank you to the sponsors of the International Cohorts Summit for their generous contribution and continued support!

NIH National Institutes of Health

Turning Discovery Into Health