International Cohorts Summit

MEETING SUMMARY

Duke University
JB Duke Hotel, Ballroom ABC, 230 Science Dr, Durham, North Carolina, USA

March 26-27, 2018

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 Cohorts Summit

EXECUTIVE SUMMARY

The International Cohorts Summit, hosted by the Global Genomic Medicine Collaborative (G2MC, https://g2mc.org/) at Duke University in Durham, NC on March 26-27, 2018, and sponsored by All of Us, National Institutes of Health, Medical Research Council, and Wellcome Trust, was conceived in 2015 by NIH and endorsed in June 2016 by the Heads of International Research Organizations (HIROs). The vision for success from this meeting and future collaborations is the creation of ***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.***

Cohorts represented at the Summit were selected based on 4 criteria: having 100K participants or more, not being disease-specific, having available biospecimens, and having at least the potential for longitudinal follow-up of participants.

Approximately 100 investigators from 24 countries representing 60 cohort studies attended and represented greater than 25 million people at their current sizes and greater than 36 million based on future recruitment targets, some with available data from the 1960s. From a pre-meeting survey, the majority of the represented cohorts have samples available, including DNA and genotyping, and almost half have whole genomic/exome sequencing data on at least some samples. Most indicated willingness to share data with appropriate patient consent, and agreed that the benefits of data sharing (enable increased cohort size, statistical power, associations, effect augmentation, and the advancement of scientific/medical knowledge/research and foster collaborations and new approaches/ideas) are understandably offset by the challenges of costs, regulations, and data harmonization.

Drs. Francis Collins and Jeremy Farrar spoke to the importance of combining large-scale cohort programs to encourage data sharing and pooling to improve scientific discovery, improve efficiencies, and maximize investments on a global scale. They emphasized that the scientific community has a responsibility to partner and share, to lead by working across borders and at a global level.

The five objectives for this Summit were:

- Improve prospects for harmonization of data, data formats, phenotype measures, consent, etc.
- Promote data and specimen sharing, and open access policies
- Examine the potential for a collaborative (global) sequencing project
- Explore the feasibility of a searchable on-line global registry of large-scale cohorts
- Create a vision: *Where do we want to be in ten years?*

The Summit was organized into 8 sessions to allow for the staging of the issues to be discussed during the breakout working sessions, and subsequent reports and identification of next steps. **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. **Sessions 2 and 3** discussed the opportunities for collaboration in the broad areas of phenotype and outcomes data, biospecimen collection, genomic and other -omic information, environmental and nutritional information, and multi-ethnic data. **Session 4** addressed data standards and privacy and **Session 6** provided an overview of the EU experience in assembling “cohorts of cohorts.” **Sessions 5 and 7** were devoted to break-out groups in the following areas:

Group 1: Creating a standardized database and registry
Group 2: IT considerations for enabling coordination, communication, centralization
Group 3: Scientific agenda with short- and long-term goals
Group 4: Policy agenda to facilitate and optimize impact of assembling cohorts
Group 5: Developing a collaborative genomic sequencing (and other -omics?) strategy
Group 6: Translation/clinical impact.

Session 8 concluded the Summit with a Summary of the Break Out group report-outs (See session 8, page 15 of Meeting Summary) and outline of possible outcomes as articulated by Drs. Collins and Farrar.

Drs. Collins and Farrar emphasized the enormous potential for great benefit to the general population and desire to assist with realizing the possible outcomes. They encouraged the exploration of opportunities and synergies with funding outside of NIH, Wellcome Trust, and other existing funders, and suggested a model of providing support for cohorts from within each country, similar to the genome project model and national infrastructure. Emphasis was placed on investment for the long term (decades) with periodic looks (every few years) to ensure appropriate productivity, timeliness, leadership and governance. They encouraged sharing best practices across cohorts to enable more effective approaches world-wide, rather than imposing a single unifying structure, and highlighted that joining a consortium could have significant benefits to individual cohorts, which will have much stronger voices with the power of this community behind them than if speaking alone.

Summit attendees articulated an initial set of compelling scientific questions (Breakout Report Summary Slide 15; also see page 19 of Meeting Summary) that would be addressable through the access to millions of individuals, such as investigating rare conditions and genotypes, enabling consanguinity and founder population studies, addressing bottlenecks with new technology development, and initiating novel pilot studies. The summit attendees recognized that this desired global platform however would require funding and other resources to address several initial desirable goals: registration and data deposition, review and compliance of country-specific data access policies, ensuring consent or re-consent processes, sequencing/genotyping support, and support for open-source data platforms and analysis platforms.

The next steps and possible outcomes from this Summit include:

- Creation of a searchable registry to facilitate collaboration across the cohorts– initially "members" vs broader global scientific community
- Establishment of foundational principles for creating consortium of cohorts (CofC) and agreement to further explore creating it
- Identification of potential key work streams to create a foundation for a possible CofC
- Creation of an organizational entity to support exploratory activities– likely G2MC and GA4GH partnership
- Outreach to cohorts not in attendance
- White paper of opportunities and challenges
- Follow-up working groups, second summit to be planned in China as offered

INTERNATIONAL COHORTS SUMMIT

Hosted by the Global Genomic Medicine Collaborative (G2MC)

Duke University
JB Duke Hotel, Ballroom ABC, 230 Science Dr, Durham, North Carolina, USA
March 26-27, 2018

MINUTES

Day 1: March 26, 2018, 8:30 AM – 6:00 PM

SESSION 1 – INTRODUCTION AND BACKGROUND
CHAIRS: GEOFFREY GINSBURG & TERI MANOLIO
8:30 – 10:20 AM

Welcome and Introductions – Geoffrey Ginsburg & Teri Manolio
Drs. Ginsburg and Manolio welcomed all participants to the International Cohorts Summit (ICS), and thanked the sponsors All of Us, National Institutes of Health, Medical Research Council, and Wellcome Trust for their generous support and the Global Genomic Medicine Collaborative (G2MC) for hosting the meeting. They both stressed the importance of the upcoming conversations and potential for enhanced collaborations among the assembled global leaders of cohorts over the next two days.

Welcome from Chancellor for Health Affairs, Duke University – Eugene Washington
Dr. Washington thanked ICS for choosing Duke as the location for this significant occasion. He described several exemplary Duke programs across mission areas, remarking they were primarily established through supportive partnerships. Noting the extraordinarily accomplished people and organizations represented in the room, he believed this Summit had an enormous potential for great benefit to the communities and populations they serve and wished the group great success in the translation of discoveries to improved care.

Vision for summit – Francis Collins & Jeremy Farrar
The notion of this Summit began in 2015 when NIH began to compile information on large cohort programs (≥100,000 participants) and these results were discussed at June 2016 Heads of International Research Organizations (HIROs) meeting. There it was agreed there was a need to bring these cohorts together. The G2MC was commissioned to organize this Summit. Cohorts were invited that met four criteria: those with 100K participants or more, not selected based on a specific disease, and had both biospecimens available and a potential for longitudinal follow-up. Some leeway was allowed for compelling reasons, such as for rare or underserved populations.

Dr. Collins stressed the importance of the combination of large-scale cohort programs. The value has been clearly demonstrated in the literature, as shown in examples of studies of blood pressure, genome wide associations, and rare genotypes (such as 'human knockouts'). This collaboration will encourage data sharing and pooling for improved scientific discovery, improve efficiencies, and maximize

investments on a global scale. Some applicable goals that could be achieved include the great potential to look at heterogeneity among populations, accelerate precision medicine, study environmental exposures worldwide, and conduct complementary studies for replication. The world is waiting for these kinds of results that can be derived from large cohorts and databases.

This Summit should strive toward the following five early objectives:

- Improve prospects for compatibility of instruments, data formats, phenotype measures, consent, etc.
- Promote data and specimen sharing, open access policies
- Examine potentials for a collaborative sequencing project
- Explore feasibility of a searchable on-line global registry of large-scale cohorts
- Create vision: *Where do we want to be in ten years?*

Dr. Collins concluded by lauding both Professor Sir Richard Peto's attendance at the Summit and his pioneering contributions to science.

Dr. Farrar articulated that science has a responsibility to partner and share, to lead by working across borders and at a global level. He believed the assembled group could realize great things for humanity, recalling the vision and success of Francis Collins and John Sulston in putting together the human genome and insisting that the science was shared and the data used to help improve health for people all over the world (and not just wealthy countries). To start this movement of sorts, he noted it is critical to effectively communicate both inside and outside the group, to be inclusive, to be collaborative, and to be global.

Summary of cohorts in attendance – Teri Manolio

A pre-event survey was conducted and results were made available in the meeting booklet. Approximately 100 attendees from 24 countries representing 60 cohort studies were present. The cohorts represent greater than 25 million people at their current sizes and greater than 36 million based on collective targets. Data available ranged from the 1960s to the present, and a majority of the cohorts have samples available, including DNA and genotyping. Almost half have whole genomic/exome sequencing data on at least some samples. Most of the cohorts indicated they had patient consent to share data beyond the initial study investigators and that they were willing to share data, albeit some may have limitations or restrictions. Most attendees believed data sharing would enable increased cohort size, statistical power, associations, effect augmentation, and would advance scientific/medical knowledge/research and foster collaborations and new approaches/ideas. Many believed the challenges to this process would include costs, regulations, and data harmonization.

Value and challenges of combining large cohorts – Rory Collins

Large prospective studies can identify the effects of complex traits, limit confounding by other factors, and determine effects of an exposure (e.g. smoking) on many different diseases. However, prospective cohorts need to be large since only a fraction of the participants develop any particular disease during prolonged follow-up. Prolonged follow-up can be particularly valuable, as shown in the Million Women dementia study where true causal relationships were revealed by excluding the first ten years to avoid "reverse causal" bias. It is important to conduct repeat measures in a subset of participants in order to be able to correct for 'regression dilution' bias (a necessity in all such observational studies), where single measures made at "baseline" tend to underestimate the real associations of disease risk with long-term "usual" levels of such risk factors. Large collections in diverse populations allow discovery of

less common diseases/genotypes than in one population. Heterogeneity of populations and exposures, rather than representativeness, can allow the assessment of the effects of different risk factors across the full range of relevant exposure levels.

Some of the challenges in large cohorts include the lack of established processes for large-scale health outcome phenotyping, constraints on access to usable data (insufficient specificity, data complexity, and sample depletion), and large costs (e.g. for cohort-wide assays that facilitate data sharing and minimize sample depletion). Strategies for improving data utility include improving interoperability of data and analysis tools (such as moving analysis to data sets), using open source data platforms, and aligning data standards to encourage sharing.

Opportunities to enhance translation for discovery to health – Geoffrey Ginsburg
Dr. Ginsburg noted the many advantages large cohort studies can have on discovery, translation, and delivery of knowledge. Risk and resilience factors can be determined. Novel predictive models of health and disease can be developed. Algorithmic approaches to prognosis and diagnostics can be refined and drive down costs. New devices/monitors and mobile phone technology (>80% of people now have one) can provide decentralized data collection. Large cohorts can make integration of personal, clinical, and biological information possible; provide continuously updated estimates of individual risk and health behaviors of populations; and provide a more profound understanding of health and disease to inform development of new therapeutics and diagnostics. Although we may be at a stage of discovery and clinical confirmation, there are still many steps ahead toward giving back to the community via implementing derived knowledge into clinical care.

Discussion
Return of research results was discussed. It is not always understood and/or desired by participants or physicians, and false positives have led to clinical procedures associated with risks. Return should be tailored to the individual and their provider and that rolling out in a phased way works well—first sharing most easily-understood results and then moving to more complex information such as polygenic risk works well. The Nurses' Health Study provides annual updates/newsletters on how samples/data have been used and that aggregate data can still be highly valuable. Privacy and time spent, rather than return of results, seemed to be bigger concerns to participants.

Session 2 – Opportunities for Collaboration Across Cohorts
Chairs: Nicola Mulder & Rory Collins
10:40 AM – 12:45 PM

Obtaining phenotype and outcome data from electronic health records and digital platforms – Josh Denny (US), Cathie Sudlow (UK), and Zhengming Chen (Asia)

- Dr. Denny discussed some of his work with the Electronic Medical Records and Genomic Research (eMERGE) Network and the All of Us Research Program. Through machine learning algorithmic analysis and data capture from electronics health records (EHRs), a 'phenotypic risk score' was developed in eMERGE. Several algorithms were demonstrated including hypertension, hypothyroidism, and drug response as well as replications the group has created for GWAS associations. Early work is being done in All of Us by aggregating data from both healthcare settings and from individuals themselves and aligning these with common data standards.

- Dr. Sudlow described some of the work underway at the UK Biobank. The Biobank contains data on 500K participants aged 40-69 years recruited in 2006-10, including extensive baseline questions and physical measures, stored biological samples, and data on 100K participants using portable wearable devices. They are working on collecting 100K participants with multimodal imaging (22K now). She demonstrated some of their work in dementia using the positive predictive value of routine healthcare data. Implementing multiple sources of unstructured data sources to enable deep phenotyping, and obtaining this at a national scale, present considerable challenges.

- Dr. Chen discussed the China Kadoorie Biobank's (CKB) efforts. The CKB has >512K participants tracked indefinitely via electronic record linkage with banked samples for long-term storage. Periodic resurvey of 5% surviving participants and an expert adjudication portal for select diseases (with over 70K completed) are ongoing. Future work for disease phenotyping includes data standardizing, ICD-10 coding of new events, and extending adjudication to other disease areas.

Value of biospecimen collection & biobanking – Nancy Pedersen
Dr. Pedersen presented via webinar the value of collecting biospecimens as well as the challenges and opportunities she has experienced. LifeGene maintains 1.5M aliquots and strives toward getting sample collection 'right'. There are many issues to consider including: storage issues, time in freezer, freeze thaw issues, and sample depletion. She has advocated for LifeGene to be standardizing collection of biospecimens, integrating with the Nordic countries national registries, collecting specimens longitudinally, and saving 10-20% of samples for future access after 10-20 years.

Value of genomic information and how to gather it – Matt Nelson
Drugs with human genetic evidence are twofold more likely to be brought to market successfully. Sources of genetic information for pharma trials have historically been public databases and collaborators, with limited information coming from pharma clinical trials. Emerging sources include cross-biobank research and other EHR-linked data. He could not overstate the value GSK has seen in establishing a link with UK Biobank. Some diseases require 5-10M patients to get the appropriate power with gene expression. Cost can have considerable impact on scale of genotyping versus sequencing: with $10M USD, given current test costs, they can perform 200K genotype arrays, 33K exome sequencing, or 17K genome sequencing. Sequencing in consanguineous populations identifies disproportionately more knock-outs. Advances in genetics research driven both by technology and scientific culture, such as exhibited at this Summit, will propel scientific advances.

Value of other -omic information and how to gather it – John Danesh
Dr. Danesh discussed the potential value of molecular 'omics' as a way to offer new insights into biology, disease aetiology/subclassification, risk prediction, and therapeutic targeting. Omics should be viewed in terms of context-specific effects, recognizing that they have considerable added complexity over DNA sequence data as they're dynamic rather than static and are tissue specific. They're also quite diverse in the research questions they address. There are some challenges to validate assays in population studies and understand complexities in interpreting assays. They have worked with the SomaScan assay in the INTERVAL study of 150,000 blood donors in England, where abundance of 8,000 proteins in over 200 cell types varied across 8 orders of magnitude. Predictive inference is enabled by validating these assays on a population scale by plotting genetic variants in relevant genes across protein levels. Plasma proteomics has the potential to yield potential causal insight into disease, recognizing that discovery power can be substantially enhanced after correction for non-biological variation. Overall, multi-omics can help address the post-GWAS grand challenge of bridging molecular gaps from genotype to disease

and 'omic' assays have common and assay-specific technical and interpretive challenges. Several assays are being used at population scale, with results being pooled across cohorts. Two cohorts to add are the UK Blood Donor cohort (150K participants) and the South Asian Cohort (95K participants).

Discussion

- How to validate results in EHR systems, particularly at global scale? Some algorithms may transfer well across systems, but different coding is used across different national systems (e.g. ICD-9 versus -10, -11 and SNOMED, etc.). Commonalities will exist, at least in terms of approach, mapping through ICD codes, etc. At CKB, they have already reduced heterogeneity across 10 different sites. With proper planning and standardization, this is doable. Efforts to do this retrospectively were seen as challenging; prospective implementation is more feasible.

- A potentially efficient approach to questionnaire harmonization would be to apply natural language processing (NLP) to studies' instruments to extract comparable information, as is done with EHRs.

- Close partnerships with assay vendors are essential not only to secure competitive prices but more importantly to ensure assay implementation and interpretation are appropriately adapted to work at very large scale.

- Biorepository management should enable use by multiple participating groups. However, there is value in both patience and procrastination; i.e., wait until a test can be afforded, wait for new kinds of tests/assays to be developed, and plan bigger picture projects rather than one-offs that may deplete samples for marginal gains. If it were possible to create massive resource of -omics data across all cohorts, the value would be unprecedented. Investment in biobanks, including not only genomics infrastructure, but phenomics infrastructure, should be viewed as a government's responsibility to its people similar to physical infrastructure. The cohorts should continue to push for scientifically valuable early wins, yet should be focused more on preaching patience to funders for grander, long-term discoveries.

SESSION 3 – OPPORTUNITIES FOR COLLABORATION ACROSS COHORTS (CONT'D)
CHAIRS: FRANCIS COLLINS & JEREMY FARRAR
1:45 PM – 3:10 PM

Value of environmental information and how to gather it – David Hunter

Dr. Hunter discussed how differences in rates of most diseases between countries and over time within countries are due to differences in environmental and lifestyle risk factors, rather than genetic differences. Differences in weight, alcohol consumption, air quality, and physical activity are major drivers in the burden of global disease. With some exceptions (e.g. drug idiosyncrasies) few supra-multiplicative gene-environment interactions have been found—genetic and environmental and lifestyle risk factors appear to be independent and the risks are multiplicative. There are always inaccuracies in self-reported data; however, Dr. Hunter showed examples in which prospectively-collected self-reported data on weight, alcohol consumption, and physical activity gave essentially the same results as objectively collected data or results from Mendelian randomization studies. Ideally, data collection should be standardized across the large cohorts—for example, the NHGRI-funded PhenX project offers standardized questionnaires available in multiple languages. Larger sample sizes are needed to analyze less common diseases prospectively, and these can be best obtained by combining self-reported data

collected on a larger scale, geolocation data, clinical and outcome data from interoperable EHRs, and system-wide outcome coding.

Value of nutritional information and how to gather it – Walter Willett
Dr. Willett discussed the role nutrition can play in large scale cohorts, as was demonstrated in their Nurses' Health Study. The high burden of coronary heart disease and most non-communicable disease is due to dietary and other non-genetic factors; e.g., upwards of 92% of Type 2 Diabetes may be prevented by diet and lifestyle. Assessment of diet, physical activity, and adiposity is essential in cohort studies, particularly when including -omics analyses. Studies need to be of long duration for many disease endpoints. Childhood and adolescent exposures may be particularly important, and repeated measures of these exposures are needed. The best single measure of diet is usually a well-designed food frequency questionnaire. For weight and height, self-report can be remarkably valid in high-income countries. Cohort studies, combined with short-term feeding studies with risk factors as endpoints, will usually provide the best evidence for translation to policy and recommendations.

Working with multi-ethnic data – Sekar Kathiresan
Dr. Kathiresan discussed the need to find mutations that protect against disease and to develop medicines that mimic them. Null mutations can provide a clear direction of effect, and null mutations that reduce risk are particularly useful for therapeutic target selection. For example, the PROMIS cohort studying consanguineous families in Pakistan has identified the world's first APOC3 homozygous null individuals. The APOC3 mutation may protect from heart attack by producing much lower blood triglyceride levels, and this discovery of a human 'knockout' allows new medicines to be developed to mimic this genetic effect. Large-scale cohorts make discovery more likely, particularly when the greatest number of unique populations across the world are included. Identifying patients through a polygenic risk model may also lend itself to creating a polygenic score, as was shown with MI, via a genome-wide set of SNPs to identify individuals with risk equivalent to a monogenic mutation.

Discussion

- Do polygenic risk scores for MI work across different ethnic groups? They do, yet ability for validation in external data sets needs to be proven for each disease. QT intervals are another trait where polygenic scores add value to monogenic variants, which can be especially useful in managing families. Studies are needed of relatives of probands who don't carry their proband's variants.

- Are nurses, as in the Nurses' Health Study, considered more reliable reporters of diet? Generally yes, but similar results have been found in other studies. Diet is difficult to track beyond two weeks for most self-reporting surveys and any new methods need to be rigorously tested.

- What is needed in terms of innovation and technology to achieve true individual-level environmental information gathering? Identifying critical bottlenecks may help drive technologic innovation, such as pattern recognition software to quantify diet components. Methods for measuring ambient air pollution may not be achieving trustworthy measurable results (existing experiments may be wrong by an order of magnitude). There is hope that new devices will become available (e.g. smart shirts to measure exposure to pollutants) in the near future. Other determinants for geographic disease rates may include natural causes or even domestic air pollution (e.g. homes with high wood smoke exposure indoors). "Experiments of nature" could be identified in very large cohorts, such as people with high PM 2.5 exposure who don't develop disease.

- How will a first-pass triage of sorts help identify human knockouts? There may be a need for advances in basic science first to understand the function of many individual genes. Large cohorts are likely the best method for finding these rare individuals, followed by validation in large datasets.

SESSION 4 – DATA STANDARDS AND PRIVACY
CHAIRS: STEPHANIE DEVANEY & GEOFFREY GINSBURG
3:30 PM – 5:00 PM

Data standards and global variant databases – **Thomas Keane**
Dr. Keane noted data often cannot be moved because barriers may be too large due to size of files, regulatory restrictions, and national legal systems. The Global Alliance for Genomics and Health (GA4GH) is aiming to produce standards to enable genomic data sharing that will allow analysis to be sent to the data where they live. The core mantra for GA4GH is to standardize on interfaces and compete on implementations, with potential implementations including better, faster algorithms for analyzing data and better ways to store data. Under the new GA4GH structure, real world genomic projects (called 'Driver Projects') give input and directly drive development of standards to meet their immediate needs for data sharing. Data access may be controlled though use restrictions by the patient and a research ID workstation vetting process.

Navigating Differences to Achieve Common Goals – **Laura Lyman Rodriguez**
Dr. Rodriguez modified the scope of this presentation from the original title 'Informed consent, data privacy'. Risks are not static or always quantifiable and, therefore, respecting consent is fundamental, but not sufficient. Privacy is not absolute, but paramount, and data security procedures will not be perfect, so they must be responsive. There is a need to accommodate differences in values, risk tolerance, privacy perspectives, and cultures. Building and sustaining trust both with patients and between organizations is paramount. There are existing frameworks to help address privacy, security, consent, such as the U.S. Precision Medicine Initiative Privacy & Trust Principles and the GA4GH Framework for Responsible Sharing of Genomic and Health-Related Data (available in a dozen languages). The best path forward is transparent, responsible collaboration within the scientific community, and adaptive policy and governance structures, which include clear accountability so as to earn and sustain trust among patients and the public.

Quantitative science to optimize the value of cohort data – **Robert Califf**
Dr. Califf noted that we seem to be entering into the fourth industrial revolution: the digital revolution, characterized by a fusion of technologies blurring the lines among physical, digital, and biological spheres. Quantitative science plays a big role in meeting society's demands. Three approaches to achieve these goals include: 1) through a combination of clinical/epi expertise and quantitative methods, considerable effort needs to go into organizing and curating the data, especially the clinical, behavioral social and environmental data; 2) the right teams of researchers are needed to enable the right data and best analysis for the specific question or purpose, and 3) the field must step up its approach to translating findings into truthful, understandable information. Both ethics and data science are too important to be done in isolation; there is a need for multidisciplinary teams together to work on these challenges. Large cohorts can help fill this need by allowing the gathering of collective intelligence and greater access for individuals.

Discussion

- Academia's incentives, such as being the first to publish, often work against collaborative efforts.
- There is a potential for harmful labelling of communities such as isolated or founder populations using genomic information.
- Concerns about data sharing may diminish when patients are unhealthy; sick people are more likely to want their data shared. In some countries where 'opt out' policies are in practice, such as France, participants seem to be more likely to want to share their data and are fine that their data will be shared with other researchers. With greater familiarity with research participation, as in Estonia where their program has been operating for almost 20 years, there seems to be greater enthusiasm for continued patient participation.
- As noted earlier, governments should feel responsible for developing cohorts as they would other vital infrastructures. It was generally agreed that anonymized data should be available to all.

Session 5 – Break-out Sessions
5:00 – 6:00 PM

- **Group 1: Creating a standardized database and registry—pros, cons, how best to do it**
  Chairs: Daniel MacArthur and Joyce Tung
- **Group 2: IT considerations for enabling coordination, communication, centralization**
  Chairs: Teresa Zayas Cabán and Thomas Keane
- **Group 3: Scientific agenda with short- and long-term goals**
  Chairs: Rory Collins and Patrick Tan
- **Group 4: Policy agenda to facilitate and optimize impact of assembling these cohorts**
  Chairs: Gad Rennert and Laura Rodriguez
- **Group 5: Developing a collaborative genomic sequencing strategy**
  John Danesh and Hakon Hakonarson
- **Group 6: Translation / clinical impact**
  Chairs: Eric Green and Dan Roden

Day 2: March 27, 2018, 8:00 AM – 2:15 PM

Session 6 – EU Cohorts and Break Out Working Sessions
Chairs: Andres Metspalu & Teri Manolio
8:15 – 10:00 AM

The EU Experience in Assembling Cohorts of Cohorts – Philippe Cupers

Dr. Cupers noted that the European Union framework programmes for research and innovation have devoted substantial funds toward building cohorts, as well as cohorts of cohorts. The EU has not funded many global population cohorts, focusing more on disease cohorts. He reviewed seven large cohorts of cohorts projects thus far supported, noting three are geared toward the elderly, one toward birth cohorts, one on children with congenital anomalies, one on children and air pollution, and one on GWAS in chronic diseases. Those projects have worked through many difficulties encountered in integrating cohorts, including harmonization of protocols, data access, funding and sustainability measures of long-term outcomes, and definition and validation of variables. Finally, he mentioned that the EU launched a

new research coordination topic on *"Building international efforts on population and patient cohorts"* that should allow to establish a strategy for the development of the next generation of integrated cohorts.

General discussion points around the EU experience include:

- 25% of all cancer deaths in EU are still due to smoking, and this fraction is still steeply rising in women. In his presentation, Philippe Cupers mentioned that the EU Horizon 2020 includes an important research priority on studying environmental factors (including lifestyle) on health conditions and risk factors.

- The EU General Data Protection Regulation (GDPR) sets up new standards in terms of data protection. The GDPR comes timely to bring a new set of "digital rights" for EU citizens in an age of increased economic value of personal data in the digital economy. This regulation becomes enforceable from 25 May 2018.

- What are the EU's lessons learned from the challenges in bringing cohorts together? What can we take forward, to avoid repeating mistakes? The EU new research coordination topic on *"Building international efforts on population and patient cohorts"* will allow to collect data and leave scientists to build on lessons learned and submit ideas on how to be most efficient in building cohorts of cohorts. Some comments were that asking scientists who are hoping for continued funding to discuss their difficulties might be problematic; they might only discuss successes and good results. There are major challenges in large scale cohorts that cohorts themselves cannot solve alone: 'it takes a village'.

Welcome from the Dean of Medicine, Duke University – **Mary Klotman**
Dr. Klotman thanked the Summit for choosing to take place at Duke University. She described several ways Duke was gearing up for the 'digital revolution', including the recent creation of a Department of Population Health Sciences. She commended the group for their sustained individual efforts and drive to come together as one, remarking that although even bigger challenges lay ahead to meld the cohorts, she believed the group would perform extraordinary feats.

Break-out Group Working sessions (reconvene)

SESSION 7 – BREAK-OUT REPORTS AND DISCUSSION
CHAIRS: CAMILLA STOLTENBERG & RORY COLLINS
10:20 – 12:20

Group 1: Standardized database – **Daniel MacArthur and Joyce Tung**

- *Creating a standardized database and registry*: The Group reviewed use cases from the perspectives of the researcher and the cohort, and proposed a four-tiered structure to gather data on cohorts: 1) cohort description; 2) data description; 3) counts and other summary data; and 4) individual-level data. The Group recommended the ICS should periodically surface what is available in each cohort, promote collaboration, and leverage existing metadata collections/questions rather than duplicating effort, while each cohort contributes up to the tier they choose. The end result will be a queryable platform, rather than database, in the cloud.

- *What challenges need to be addressed to optimize the value of sharing information?* The Group strongly felt tasks such as fielding inquiries, doing analyses to evaluate feasibility, and sharing datasets, all take real work and are not free. A stable flow of resources will need to be secured. Some processes may be automated, yet many will need real human effort. Data sharing models will vary by cohort (e.g., 1:1 relationship building vs. limited interaction) and infrastructure may require considerable technical work. Technology companies may be able to help here, but data should not be held centrally by these companies.

- *How might we move towards standardization and sharing of individual-level data across cohorts*? Storing individual-level data poses far greater challenges than storing cohort metadata, including hard limits on data leaving particular countries. Particular difficulties are seen in phenotype harmonization. Governance via an independent global federation seems the best option.

- *What types of projects could this registry facilitate?* One possibility was to create a "reproducibility network" to rapidly validate associations discovered in a single cohort. The cohorts could also focus on rare diseases or rare exposures that require massive sample sizes to study and/or conduct studies that take advantage of genetic and environmental diversity across cohorts. The Group foresees significant efforts needed to overcome language and translation challenges. It will be important to predetermine any work does not duplicate similar ongoing efforts such as the Cancer Consortium. An interesting paper might be written on various cohorts' approaches to data sharing.

Group 2: IT considerations for enabling coordination, communication, centralization – Teresa Zayas Cabán and Thomas Keane

- *What data are collected/available for each of your cohorts and what are their formats?* This is highly dependent upon the findings of Group 1. This Group noted a merge with Group 1 may be warranted to promote symbiotic, non-duplicative work.

- The second and third questions posed to the Group also reinforces the need to merge efforts with Group 1: *What are sources of those data*? and *How are data collected in each of your cohorts stored?* They will need to understand how are the data distributed, accessed, and made discoverable.

- *Which model is suitable for large scale cohorts?* A federated joint cohort analysis will be needed to benefit clinicians and research discoveries. Centralization will reduce duplication of effort and allow a model with few single points for data discovery and access to obtain more sustainable funding (it may be easier to get funding for few much larger portals). Federation may allow circumvention of jurisdictional restrictions on export, particularly healthcare data, and promote the creation of a data safe haven, with a meta-data dictionary, meeting enhanced security requirements. This would not be trailblazing either, as Pharma already works in such a model.

  Standard interfaces will be required (e.g. - UK, DataShield) which will allow users to request access, login to another infrastructure, analyze data, and export allowed results. The key is to avoid having to write a specific pipeline to run analysis at every site anew. The ICS can provide a standard set of analysis modules to choose from at each cohort. Authorization and access can be done through 'researcher library cards' to vet bona fide researchers. It was noted proving bona fide standards could be problematic and may be exclusionary. The tools already exist for cohort data harmonization (ontology mapping tools exist) and integration (e.g. - CDISC format), yet these may be

some of the biggest challenges to ICS. Discoverability, by genotype (Beacon), phenotype (Matchmaker Exchange), and by data use (data use ontology), may promote hypothesis-free research. Imaging data may be difficult to integrate, while cell images may prove 'identifiable'.

It was discussed that the biggest challenge to ICS may actually be the cohort work itself. Understanding just one cohort takes an incredible amount of time and effort. This can produce superficial science, not good science. There needs to be a distinction between data taking and data sharing. It also may prove likely a shared analyses will need to involve those who generated the cohort data in the original studies to truly understand the data. It was agreed a happy medium that does not require the people who created the cohort to be involved in every single analysis or use of the data would be preferred; the ICS should aim to create tools to facilitate better use of the data by all comers. These efforts should leverage extant work in data models, data sharing, and data standards from other research data networks such as PCORNet.

Group 3: Scientific agenda with short- and long-term goals – Rory Collins

- *What enhancements to existing cohorts would most increase their utility and promote data sharing?* The ICS should promote collection of biological samples in existing cohorts, including new samples to show change over time as well as newer sample types (microbiome, RNA). The ICS could develop population-specific genotyping arrays, which may require some WGS to inform array design and imputation. Cohorts should standardize and characterize novel –omic assay methods (including data processing, analysis and reporting) and support cohort-wide genotyping and other –omic assays. Phenotyping of health outcomes should also be standardized, using algorithms based on health record systems and other sources (including access to tumor samples). Repeated measures are critical for assessing change. The combined cohorts could work on developing novel methods for charactering exposures (imaging, environment/socioeconomic data) and outcomes (mobile technology for cognitive decline, arrhythmias) and generic data visualization and management systems to support use and sharing of individual cohorts.

- *What are the highest priority scientific questions that could be addressed by a cohort of cohorts?* The ICS could provide support for collaborations among cohorts to address specific scientific questions by combined analyses of the data, produce context-specific analyses of the local relevance of risk factors that could better inform global burden of disease and other estimates, and assess what determines "health" in different settings. One strength could be the power to address questions that are related to conditions for which there are likely to be too few cases or individuals of interest (e.g. young people; ethnic groups) in any individual cohort.

  The ICS could support the development of systems to facilitate long-term follow-up of health outcomes in lower-income populations and could also facilitate access to the widest range of health outcome data for long-term, comprehensive follow-up of all participants. Both issues are likely to require the engagement of research funders with governments while navigating all the data protection obstacles. It may be difficult to obtain outcome data or follow-up data in resource-limited settings such as some parts of Africa, yet cohorts in sub-Saharan Africa are growing as is in-depth networking, with many nations doing it themselves and not necessarily disseminating or notifying others. The types of environmental data that are a high priority need to be identified, yet environmental exposures of the past may be difficult to measure.

The cohort of cohorts would be of great value of work for young investigators and the ICS should make it a part of the scientific agenda to have a training platform and marshal resources to train the next generation of epidemiologists. Other strengths of the ICS are that it could negotiate lower prices for sequencing, gain affordable access to storage space, and develop training opportunities. One concern expressed was that smaller cohorts could get lost in a huge consortium.

Group 4: Policy agenda to facilitate and optimize impact of assembling these cohorts – Gad Rennert and Laura Lyman Rodriguez

- *Common challenges for international collaboration:* The ICS would need to demonstrate benefit to stakeholders (funders, participants, investigators, public) and show how shared cohort research (and ICS collaboration) is meeting the stated mission. Institutional interpretation of local/national regulations, heightened concerns with sample sharing, and lack of education/understanding of IRBs, providers, and public (policy makers) will need to be overcome. Established local enclaves of data access frameworks and inclusion and management of data generated in Native/Aboriginal populations may present challenges to be addressed.

- *Common needs/potential benefits for collaboration:* The ICS will need to hold joint policy and scientific design discussions for initiatives and provide a clear distillation of collaborations' guiding principles and goals. International principles will need to inform local/national decision-making while leaving the ability for cohort studies to make choices for participation intact. GA4GH has produced materials that could support some of these efforts. Funding sources will clearly be needed and communication strategies will have to be developed, including an avenue to share lessons learned and strategies to overcome barriers.

- *Actions or pilot efforts/proof of principle to consider for follow-up over next 1-3 years*: Near-term goals include creating a governance description for the inventory of cohort studies, defining cohort policy challenges (e.g., incorporating GDPR guidelines and implications), and engaging with primary funders of cohorts, which may include for-profit entities. Some early wins could help to garner interest (and funding!). A harmonized informed consent process would allow more flexibility, yet harmonized measures should not mean the lowest common denominator.

Group 5: Developing a collaborative genomic sequencing strategy – John Danesh and Hakon Hakonarson

- *What are the key questions we can only address though large-scale international collaboration?* This question was added by Group 5 to promote discussion. The ICS is well situated with both genomic and exposure diversity and could conduct both migrant and rare diseases studies. It is well-positioned to help address global problems such as obesity, alcohol-related diseases, and toxic substances exposures. A cohort of cohorts can drive sequencing and other -omics costs, could conduct large scale GWAS/PheWAS studies, and could provide an avenue for hypothesis-driven approaches.

- *What kinds of sequencing or other -omic data would be useful for individual cohorts?* Most valuable would likely be whole genome sequencing with sharing of WGS data files, while WES (identifying new variants in known GWAS genes) did not seem a high priority. At the outset the ICS could leverage existing data and share new SNP-array genotyping data across all samples; interest among non-European cohorts was greater for array genotyping supplemented by subgroup WGS.

- *What aspects of a collaborative sequencing strategy, in addition to low cost, would facilitate obtaining and sharing these data? S*ome early-term goals could include: leverage existing efforts on data file harmonization (GA4GH and others), assist low income countries needs and multiple other sites low in funding, and promote access to the limited numbers of RNAseq, proteome, epigenetics data, microbiome data, and metabolome data.

- *What methods/tools are optimal for data harmonization across different sites to address platform diversity/uniformity, batch effects and related issues?* While exomes have too much variability, imputed SNP-array data has proven to work across multiple continents. WGS data harmonization has proven effective at US sites. Sharing phenotyping data should not be problematic data and if metabolome data can be generated cost-effectively there might be value in sharing. Coordination among multiple organizations, as well as among the ICS Working Groups, would be imperative, particularly in cataloguing cohort information, responding to queries, and identifying the right cohorts. There could be two potential ways to approach harmonization: create an agnostic platform (set up procedures, rules, capabilities) or launch a science-driven grand challenge (or multiple challenges) based on specific research questions. The grand challenge approach seemed to be preferred. Perhaps ICS should focus on genotyping first, with regional arrays and an eventual global array, as an early objective.

Group 6: Translation / clinical impact – Eric Green and Dan Roden

- *What are the opportunities for translation of cohort findings to improved clinical care and population health?* Findings should advance medical practice, assist drug development, unveil generic opportunities, and add to the knowledge base of population health and policy. The Group additionally explored potential exemplar projects that could benefit each member of a cohort of cohorts. Ideas included: an international human knockout project, standardization in the implementation of return of results, and country (cohort)-specific risk prediction using standardized methodology.

- *What are the major barriers to clinical and population health translation and how can they be dealt with?* Barriers include the need for exemplar projects, while navigating variable healthcare delivery systems and addressing disparities, lack of diversity, and lack of evidence for clinical utility. Additional difficulties will likely involve regulatory roadblocks, reimbursement, and privacy regulations across countries. Thousands of people have had private genetic testing done, particularly in the US, can they be invited to become a cohort/ have their data entered in a publicly-accessible database? At 23andMe, 85% of customers consent to participate in research, including some longitudinal studies, which makes it a cohort of sorts; in addition, all their customers can download their raw genetic data and choose to upload it to any other service.

SESSION 8 – WORKING LUNCH, SUMMARY AND NEXT STEPS
12:20 – 2:15 PM

Summary, outline of 1-year plan – Geoffrey Ginsburg and Teri Manolio

Numerous action items were developed in this two day conference toward achieving the ICS vision. Drs. Manolio and Ginsburg reviewed the Summit's discussions and path moving forward. A vision for success was articulated as "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." The

summary slides are reprinted below with a few added notes (preceded by "-") from the discussion. Italics indicated points of emphasis during the discussion.

Group 1: Create a standardized database and registry

1) Start with cohort-level metadata, work toward unified database of individual-level data
2) Leverage existing metadata collections, automate as much as possible (reference NCI Cohort Consortium, UK Dementia Platform)
3) Tiered structure: each contributes at level of comfort, queryable
    a. 0: Cohort website
    b. 1: Cohort description (like program book), mechanisms for access
    c. 2: Data description (demographics, data collection instruments, -omic data)
    d. 3: Counts of phenotypes, sample types, update
    e. 4: Individual-level data
4) Funding and infrastructure needed for fielding queries, depositing data, companies may have solutions but should not hold the data
5) Projects that could be facilitated: *build reproducibility network*
6) Address data security with cloud solutions
7) Need to understand country-specific regulations on use of servers and data produced from them
8) Create scalable, transportable systems for extracting follow-up information, outcomes
9) Develop repository of SOPs for sample collection and storage, assays and protocols, assessments of validity/quality, analytical methods

Group 2: IT considerations for enabling coordination, communication, centralization

1) Need to define cohorts, what data stored and used
    – Strict cohort inclusion criteria can be relaxed for lower resource, small population, etc. cohorts
2) Federation vs centralization
    a. Centralizing reduces duplication, more sustainable funding
    b. Federated addresses jurisdictional restrictions, sheer size; sustainability of smaller portals in question
    – Federated was preferred, with caveat each cohort can maintain some FTE in the group
3) Standard interfaces needed, some extant examples
4) Interoperable analysis a goal of GA4GH workstream
5) Authorization and access: once fully authorized get expedited access
6) Operationalizing consent– funding to harmonize and standardize consent, ensure access includes consent
7) Discoverability: searchable databases, e.g. - GA4GH
8) Develop *Data Use Ontology* to tag datasets with use restrictions
9) Specifying hypotheses is not always possible, enable exploration and pattern identification through hypothesis-free machine learning
10) Challenge of legacy data
11) Measuring impact
12) Opportunity to negotiate with cloud venders
13) Highest priority: Work with Group 1 to get accurate registry, select data standard

Group 3: Scientific agenda with short- and long-term goals

1) Enhancements to existing cohorts:
    a. Collect new samples for repeated measures or novel assays

   b. Population-specific genotyping arrays
   c. Standardize novel assay types: *prioritize/choose*
   d. Support cohort-wide genotyping and other –omic assays
      – Subgroup and case/control assays may still make sense for certain analytes; no need to perform all assays in all people
   e. Standardized phenotyping approaches especially with EHRs
   f. Novel environmental measures and mobile technologies not possible through EHR linkage: *prioritize/choose*
   g. Generic data visualization methods
2) Scientific questions
   a. Good examples in past, almost always require close collaboration and independent support
   b. Determinants of health
   c. Rare conditions or subgroups
   d. Harmonize/standardize only what really matters– novel -omic assays
   e. Develop systems for long-term health outcomes in LMIC
   f. Facilitate research access to health outcome data, data protections
3) *Will require engagement of funds with governments to convince relevance of research to health care*

<u>Group 4: Policy agenda to facilitate and optimize impact of assembling these cohorts</u>

1) Challenges:
   a. Defining "what's in it" for… investigators, cohorts, countries
   b. Differing institutional interpretations of regulations
   c. Sharing samples more difficult vs. data/metadata
   d. Differing approved uses by cohorts
   e. Including native/aboriginal communities
   f. Including for-profit entities
2) Needs/Benefits: combine scientific and policy discussions
   a. Define what collaboration trying to do, high-level principles
   b. Develop/adopt international principles, build on GA4GH
   c. Develop protocol for project review– selection and participation
   d. Resource/platform to share lessons learned
3) Pilot efforts
   a. Governance description
   b. Cohort policy "traits" added to metadata
   c. Identify current collaborations, what's worked, lessons learned
   d. Understand implications of GDPR: *Engage primary funders for shared benefit and power*
4) Potential for common consent, at least for given project
5) Develop pre-competitive spaces for industry to interact with cohorts
6) Use policy frameworks for responsible sharing, obtaining consent, ensuring privacy– All of Us and GA4GH
7) Develop international strategic agenda for CoC coordination

<u>Group 5: Developing a collaborative genomic sequencing (and other -omics?) strategy</u>

1) Key questions only through large-scale collaborations
   a. Genomic and exposure diversity, *migrant studies*
   b. Rare diseases: human knock-out and homozygous deletions
   c. Drug repurposing opportunities

   d. Global problems: obesity, toxic exposures, alcohol-related diseases
   e. Microbiome across ethnicities and exposures
2) Very large projects will drive down costs
3) Centralized coordinating function enabling queries to identify most informative cohorts for specific question
4) Types of data of most value
   a. WGS, to be shared, WES not enthusiastic
   b. Leveraging existing GWAS data and SNP-array genotyping ($50-100M)
   c. Would need cohort-specific sequencing (few hundred?)
   d. Phenotyping data (no sharing issues?) and metabolomic data
5) Design collaborative sequencing strategy
   a. Leverage existing GA4GH efforts at standardization, reduce artifacts
   b. LMIC need funding, level playing field
6) Methods/tools
   a. Imputed SNP arrays rather than exomes (too much variability)
   b. WGS data file harmonization (TOPMed > 100K)
   c. May need charter or principles
7) Conceptualize as agnostic platform vs. science-driven questions
   a. Choose handful of grand challenges: genomic variation (knock-outs) and exposure variation (alcohol)
8) Genomics is easy: same in all cell types, stable
9) Many complexities of proteomic data: population validation, interpretation, biologic variation
10) Need close partnerships with developers of assays, can work iteratively to improve them: *How to decide when assay ready for millions?*
11) Large numbers reference samples in key subgroups (elderly?)

<u>Group 6: Translation and clinical impact</u>

1) Opportunities:
   a. Advance practice– diagnosis/prognosis/treatment; Mendelian, PGx, genetic risk scores
   b. Drug development
   c. Generic: health literacy, exemplars for teaching, evidence generation using large simple trials, learning healthcare systems
   d. Population health and policy: new knowledge moves to policy
2) Barriers:
   a. Variable healthcare systems, disparities, diversity, evidence
   b. Hand-off from evidence to implementation
   c. Regulatory, reimbursement, ethics, “academic territorialism”
   d. Engaging industry (real and perceived)
3) Exemplar projects
   a. Standardize implementation of RoR: Familial Hypercholesterolemia or cancer
   b. Country/cohort-specific risk prediction with standardized methods
      i. GRS and non-genetic risk factors across ancestries
4) Provide continuously updated estimates of individual risk and health behaviors of neighborhoods and populations; example of Mexico City Cohort where poor glycemic control identified as likely cause of high mortality from renal disease and other causes, leading to a rapid public health intervention in Mexico
5) Integration of personal, clinical, biological information

Compelling scientific questions addressable with millions of individuals

- Rare conditions (CKB venomous snakebite), subgroups, exposures
- Rare genotypes: human knock-out project, extremes of risk
- Consanguinity and founder population studies: for collaborations?
- Critical bottlenecks: drive technology development
- Pilot studies:
    - Utilize repository of e-phenotyping algorithms (PheKB) and test transportability across different countries' EHRs
    - Apply NLP to cohort studies' data collection instruments (or consents?) to extract data just as doing with EHRs
    - Identify high-risk individuals for early disease detection, recognize when undetected disease biasing early outcomes

Funding needs

- Register and deposit data
- Review country-specific data access policies and ensure compliance
- Harmonize consents, re-consent
- Scalable phenotyping of outcomes: ascertainment (suspected cases), confirmation (case-ness), classification (subtypes, details)
- Collaborative analyses
- Adding value to existing cohorts– cohort wide assays, novel methods
- Patience: invest for long term, avoid pushing for quick publications
- Sequencing/genotyping support in LMIC to level playing field
- Support for open-source data platforms, analysis environments, data deposition

Possible outcomes from this meeting

- Creation of a searchable registry to facilitate collaboration across the cohorts– initially "members" vs broader global scientific community
- Foundational principles for creating consortium of cohorts (CofC) and agreement to further explore creating it
- Identification of potential key work streams to create a foundation for a possible CofC
- Organizational entity to support exploratory activities– likely G2MC and GA4GH partnership
- Outreach to cohorts not in attendance
- White paper of opportunities and challenges
- Follow-up working groups, second summit to be planned

Consensus vision and path forward – **Francis Collins and Jeremy Farrar**

Drs. Collins and Farrar reiterated their thanks to the assembled members of the Summit. They agreed that there was an enormous potential for great benefit to the general population and wanted to assist with seeing the listed action items realized. Both noted that there is already substantial funding available and further funding could be available, but they wanted to stress that the Summit looks for opportunities and synergies with funding outside of NIH, Wellcome, and other existing funders. Providing support for cohorts within each country, similar to the genome project model and the necessary national infrastructure described earlier, would be a viable model. Investment needs to be for the long term (decades), with an appreciation that quick wins should not be at the expense of long term benefits. While aiming for long-term commitments it is reasonable to have periodic looks (every five

years?) to ensure appropriate productivity, timeliness, leadership and governance. We should avoid trying to impose a single unifying structure on everyone; rather, sharing what works and best practices will enable more effective approaches world-wide. Joining a consortium could have significant benefits to individual cohorts, which will have much stronger voices with the power of this community behind them than if speaking alone. Dr. Collins floated the idea that the Summit reconvene sooner than a year from now, and mentioned that a site in China has offered to be the next place to meet. Both emphasized they saw this as a long-term effort and wanted to welcome more cohorts over time. In conclusion, there is strength in numbers and together we have the numbers!

ATTENDEES

NAME	AFFILIATION	COUNTRY
Ada Al-Qunaibet	Saudi National Biobank	Saudi Arabia
Jesus Alegre Diaz	Mexico City Prospective Study	Mexico
Fowzan S Alkuraya	Saudi Human Genome Program	Saudi Arabia
Garnet Anderson	Women's Health Initiative (WHI)	USA
Cori Bargmann	Chan-Zuckerberg Initiative	USA
Valerie Beral	Million Women Study	UK
Robert Califf	Duke University	USA
Juan Pablo Casas	UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS)	USA
Philippe Cupers	European Commission	Belgium
Mark Caulfield	Genomics England / 100,000 Genomes Project	UK
Zhengming Chen	China Kadoorie Biobank	UK & China
Justina Chung	Global Genomic Medicine Collaborative (G2MC) Global Alliance for Genomics and Health (GA4GH)	Canada
Francis Collins	National Institutes of Health (NIH)	USA
Rory Collins	UK Biobank	UK
John Danesh	University of Cambridge	UK
Joshua Denny	Vanderbilt University	USA
Stephanie Devaney	U.S. All of Us Research Program	USA
Rajesh Dikshit (via teleconference)	Barshi Cohort Tata Memorial Centre	India
Lena Dolman	Global Genomic Medicine Collaborative (G2MC) Global Alliance for Genomics and Health (GA4GH)	Canada
Robert Eiss	National Institutes of Health (NIH)	USA
Jonathan Emberson	Mexico City Prospective Study	Mexico
Arash Etemadi	Golestan Cohort Study & Persian Cohort Study	USA
Jeremy Farrar	Wellcome Trust	UK
Neal Freedman	Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO)	USA
John Gallacher	Medical Research Council (MRC) Dementia Platform	UK
Susan Gapstur	Cancer Prevention Study II (CPS-II)	USA
J. Michael Gaziano	Million Veteran Program	USA
Matthew Gillman	Environmental influences on Child Health Outcomes (ECHO) Cohort	USA
Geoff Ginsburg	Duke University Medical Center Global Genomic Medicine Collaborative (G2MC)	USA
Roger Glass	Fogarty International Center	USA
Marcel Goldberg	Constances Project	France
Peter Goodhand	Global Alliance for Genomics and Health (GA4GH)	Canada

Eric Green	National Human Genome Research Institute (NHGRI)	USA
Francine Grodstein	Nurses' Health Study (NHS), NCI	USA
Jeremy Grushcow	Newfoundland and Labrador Genome Project Sequence Bio	Canada
Carolina Haefliger	AstraZeneca Integrated Genomics Initiative	Sweden
Christopher Haiman	Multiethnic Cohort Study	USA
Hakon Hakonarson	Children's Hospital of Philadelphia (CHOP) Biorepository	USA
Arthur Holden	Genomic Resources Consortium, Ltd.	US
David Hunter	Harvard University	USA
Rahman Jamal	Malaysian Cohort Study	Malaysia
Sun Ha Jee	Korean Cancer Prevention Study (KCPS)	Korea
Jae-Pil Jeon	Korea Biobank Project	Korea
Lixin Jiang	China PEACE (Patient-centered Evaluative Assessment of Cardiac Events) Million Persons Project	China
Mattias Johansson	European Prospective Investigation into Cancer and Nutrition (EPIC)	France
Farin Kamangar	Golestan Cohort Study & Persian Cohort Study	USA
Sekar Kathiresan	Center for Genomic Medicine, Massachusetts General Hospital	USA
Thomas Keane	European Bioinformatics Institute (EMBL-EBI)	UK
Sung Soo Kim	Korean Genome and Epidemiological Study (KoGES) & Korea National Institute of Health	Korea
Mary Klotman	Duke University	USA
Paulo Lotufo	ELSA-Brazil	Brazil
Daniel MacArthur	Broad Institute of MIT and Harvard	USA
Ytina Mangum	Duke University	USA
Teri Manolio	Division of Genomic Medicine National Human Genome Research Institute (NHGRI)	USA
Koichi Matsuda	Biobank Japan	Japan
Joe McNamara	Medical Research Council (MRC)	UK
Martin McNamara	45 and Up Study	Australia
Mads Melbye	Danish National Biobank	Denmark
Beatrice Melin	Northern Sweden Health and Disease Study	Sweden
Andres Metspalu	Estonian Genome Project Estonian Genome Center of University of Tartu	Estonia
Nicola Mulder	H3Africa and H3ABioNet	South Africa
Yoshinori Murakami	Biobank Japan	Japan
Michael Musty	Duke University	USA
Øyvind Næss	Norwegian Family Based Life Course Study	Norway
Matthew Nelson	GSK	UK
Thea Norman	Gates Foundation	USA
Alison Park	University College London	UK

Nancy Pedersen (via teleconference)	LifeGene	Sweden
Alexandre Pereira	ELSA-Brazil & Baependi Cohort	Brazil
Richard Peto	Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU) University of Oxford	UK
Paul Pinsky	Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO)	USA
Erica Pufall	Wellcome Trust	UK
Tejinder Rakhra-Burris	Duke University	USA
Gad Rennert	Clalit Israeli Genome Project	Israel
Gabriela Repetto	Maule Cohort / MAUCO Study Universidad del Desarrollo	Chile
Dan Roden	BioVu Vanderbilt; eMERGE Network	USA
Laura Lyman Rodriguez	National Institutes of Health (NIH)	USA
Norie Sawada	Japan Public Health Center-based Prospective Study (JPHC) and JPHC for the Next Generation (JPHC-Next)	Japan
Catherine Schaefer	Kaiser Permanente Research Program on Genes, Environment, and Health	USA
Adam Schlosser	World Economic Forum	USA
Chen-Yang Shen	Taiwan Biobank	Taiwan
Terrence Simmons	LIFEPATH (Lifecourse biological pathways underlying social differences in healthy aging)	UK
Camilla Stoltenberg	Norwegian Mother and Child Cohort Study (MoBa) & Cohort of Norway (CONOR)	Norway
Cathie Sudlow	UK Biobank	UK
Heljä-Marja Surcel	Finnish Maternity Cohort Serum Bank	Finland
Anthony Swerdlow	Generations Study	UK
Patrick Tan	Singapore National Precision Medicine Program Agency for Science Technology and Research Singapore Biomedical Research Counsil	Singapore
Joyce Tung	23andMe	USA
David van Heel	East London Genes and Health	UK
Goran Walldius	Apolipoprotein MORtality RISk study (AMORIS)	Sweden
Eugene Washington	Duke University	USA
Walter Willett	Nurses' Health Study II (NHSII), NCI	USA
Christine Williams	Ontario Health Study (OHS) & Canadian Partnership for Tomorrow Project	Canada
Marc Williams	MyCode Community Health Initiative	USA
Masayuki Yamamoto	Tohoku Medical Megabank Project	Japan
Teresa Zayas Cabán,	Office of National Coordinator (ONC) for Health Information Technology	USA
Marie Zins	Constances Project	France

SPONSORS

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

All of Us RESEARCH PROGRAM
The Future of Health Begins With You

NIH National Institutes of Health

MRC Medical Research Council

wellcome