International Common Disease Alliance

Recommendations and White Paper

July 2020

INTERNATIONAL COMMON DISEASE ALLIANCE | Maps to Mechanisms to Medicine

BLANK PAGE

Introduction

Human genetics stands at a pivotal moment.

The past several decades have seen enormous progress. Various efforts—the Human Genome Project, deep catalogs of genetic variation, to the GWAS revolution, and massive sequencing—have resulted in a wealth of knowledge about the genetics of common disease. This includes the discovery of more than 70,000 robust genetic associations to common diseases and traits, and important insights into the underlying basis of some diseases.

Yet, there is much more that needs to be done. There is again a growing sense across the human genetics community that now is the time to articulate a vision for the next phase of common complex disease genetics —to accelerate progress in moving from **Maps to Mechanisms to Medicine.**

Over the past two years, discussions among scientists across the human genetics community have led to the decision to form an International Common Disease Alliance (ICDA) as a way to engage the community.

As a first step, ICDA has developed these Recommendations & White Paper, based on the input of a diverse and engaged community of researchers, clinicians, funders, and policy experts. We have designated these documents as v1.0, because they are expected to evolve over time. We also welcome further community input on how we might continue to refine and improve these recommendations at https://www.icda.bio/.

The next step for ICDA will be to work with the scientific community to help implement these recommendations.

BLANK PAGE

Table of Contents

Recommendations v1.0

White Paper v1.0

BLANK PAGE

International Common Disease Alliance: From Maps to Mechanisms to Medicine

Recommendations v1.0

Developed by the ICDA Organizing Committee and Working Groups
(See contributor list at end)

INTERNATIONAL
COMMON DISEASE
ALLIANCE

Maps to
Mechanisms to
Medicine

BLANK PAGE

I. Introduction

Human genetics stands at a pivotal moment. The past several decades have seen enormous progress. Various efforts—the Human Genome Project, deep catalogs of genetic variation, powerful study designs and new analysis methods—have resulted in a wealth of knowledge about the genetics of common disease. This includes the discovery of more than 70,000 robust genetic associations to common diseases and traits, and important insights into the underlying basis of some diseases.

Yet, there is much more to be done. It is time to envision the next phase of human genetics—to accelerate progress in moving from **Maps to Mechanisms to Medicine (the *M2M2M Challenge*)**.

The Role of Genetics in Common Disease

For almost all common diseases, including non-communicable diseases and infectious diseases, genetics plays a substantial role in disease susceptibility, disease progression, and/or response to therapies. And, it is becoming increasingly clear that genetics offers a powerful approach to systematically propel the understanding and treatment of common diseases.

Genetics plays a unique role within medicine, because it provides a way to discover the causal biological mechanisms of any disease with no prior biological hypothesis about the cell types or processes involved. Every other biological observation associated with a disease might either be a cause or an effect. For genetic variation, the arrow of causality runs in only one direction: from genotype to phenotype.

Advances in genetics are now making it possible to comprehensively screen genetic variation in patients with common diseases to enable systematic discovery and characterization of disease variants, disease genes, disease mechanisms, and drug targets — and to develop powerful new approaches to medical care informed by genetics.

Two caveats should be stated explicitly:

(i) The Maps to Mechanisms to Medicine Challenge provides a useful framework for advancing knowledge about a disease. While advances will often flow along this linear path, we note that important information may sometimes flow backward (for example, insights about cellular mechanisms could aid in identifying target genes for non-coding associations). Also, while it will be important to develop general paradigms, we do not expect there will be a "one-size-fits-all" solution for all common diseases.

(ii) While genetics plays a key role in understanding common diseases, genetics is only part of the picture. Other factors such as environmental exposures, or social determinants of health including poverty and lack of access to education and healthcare, are also important contributors to disease. The M2M2M focus on genetics is driven by the recent availability of powerful systematic approaches that have led to increasing knowledge about genetic contributions to disease. Knowledge from genetic studies will ideally help in subsequent elucidation of non-genetic factors, and conversely.

International Common Disease Alliance

Achieving the promise of using genetics for common disease will require deep collaboration among many communities and stakeholders across academia, medicine, biopharma, tech companies, and funders.

The International Common Disease Alliance (ICDA) was formed to serve as a scientific forum to bring together these key stakeholders from around the globe in order to:

I. identify common barriers to progress across diseases and populations,
II. develop and promote collaborative solutions to overcome these barriers, and
III. convene the scientific community regularly to tackle challenges together and share results and progress.

ICDA membership is open to everyone. The Alliance is led by the ICDA Organizing Committee, which comprises 35 scientists from 14 countries on 5 continents.

In September 2019, ICDA held its inaugural scientific meeting outside Washington, D.C. to bring together participants for scientific talks and discussions. At that meeting, ICDA released a preliminary draft ICDA White Paper (v0.1) reviewing the history of the field and the challenges and opportunities ahead.

The second ICDA scientific meeting was planned for early March 2020 in Copenhagen, but was postponed in light of the COVID-19 situation then beginning to emerge in Europe. Instead, ICDA held two virtual town hall meetings.

What ICDA is — and isn't

It is important to make clear what ICDA is and isn't.

ICDA is a scientific forum comprising international stakeholders across academia, medicine, biopharma companies, tech companies, and biomedical funders.

- ICDA works to bring together the community to define current barriers to progress; identify needs and opportunities to overcome these barriers; and bring together the community to propose solutions — including through scientific initiatives and policy groups — to drive progress on the M2M2M Challenge.
- We expect that ICDA members will be actively engaged in these activities — such as piloting new experimental technologies; scaling up promising technologies to generate foundational genomic data resources; developing novel analytical methods to integrate large genetic and genomic datasets; developing new data platforms, and developing ethical frameworks for truly global and equitable collaboration.
- ICDA does not expect to decide on or fund projects. These choices should be made by funders in countries around the world — including governments, philanthropies and industry.
- ICDA will help to bring together the community to envision, stimulate, and sometimes coordinate activities across labs and countries.
- ICDA will work in partnership with existing efforts in human genetics — not seeking to duplicate the many highly functional activities already underway.

ICDA's goals are described in this "charter":

ICDA Charter

The International Common Disease Alliance will serve as a **scientific forum** bringing together international stakeholders across academia, medicine, biopharma companies, tech companies, and biomedical funders to:

- **define current barriers to progress in tackling the M2M2M challenge,** including scientific, technological, policy, computational and organizational obstacles;
- **identify needs and opportunities for new projects** to overcome these barriers in the spirit of past and present examples of public efforts and public-private projects (such as the SNP consortium, the HapMap projects, the 1000 genomes projects, the FinnGen Project, the UK Biobank, the All of Us Project, the Open Targets Initiative, and many more).
- **organize working groups to propose solutions and to drive progress**, including key knowledge, datasets, experimental technologies, computational platforms, and frameworks for data sharing and data harmonization;
- **organize scientific meetings to bring together the community** on an ongoing basis to share results, assess progress, and update plans about the genetics of common disease;
- **coordinate with funders** to ensure the work defining the barriers and proposing solutions in the white papers is of maximal utility;
- **help to facilitate international collaborations** where appropriate; and
- **undertake public communication and engagement** on issues related to common disease genetics.

Diversity, Inclusion and Ethics

ICDA is committed to diversity, inclusion and ethics. The success and sustainability of the M2M2M vision depends upon building and maintaining trust in communities around the world — especially with regard to ensuring that activities engage and benefit low and middle income countries and diverse populations.

More broadly, ethical considerations — especially with regard to the protection and benefit of patients and participants — cannot be a side activity, but must be embedded in every aspect of ICDA activities.

Developing ICDA Recommendations

Following the first ICDA meeting in September 2019, the ICDA community began work to develop concrete recommendations—intended for researchers and funders—about how to propel progress toward tackling the M2M2M Challenge. The process involved three working groups and the Organizing Committee, including ~100 participants, ~40 meetings, and tremendous collaboration.

Draft ICDA Recommendations (v0.9) were approved in early February 2020 for release as a public draft version, along with an updated version of the ICDA White Paper (v0.9). ICDA then solicited input from the wider scientific community, with the intention that the documents would be finalized at ICDA's planned meeting in Copenhagen in early March 2020. With the postponement of the Copenhagen meeting due to the COVID-19 pandemic, two virtual town halls were organized to gather further input into the recommendations. Over the subsequent weeks, the recommendations were revised in response to the input received.

The ICDA Recommendations (v1.0) presented here thus reflect the input of a wide community.

The ICDA Recommendations range broadly. They include bold scientific projects that may take 5-10 years to complete; foundational genetic and genomic resources; critical computational tools; activities to propel therapeutics; important areas for new research and training; and strategies to promote global equity.

We expect the ICDA recommendations to continue to evolve and mature as the field progresses. We welcome your input on how we might continue to refine and improve these recommendations at https://www.icda.bio/.

II. Framework for ICDA Directions and Recommendations

The ICDA Recommendations are intended to provide an **overall framework** for what it will take to meet the M2M2M Challenge. The framework consists of 8 key directions, with 23 specific recommendations. This section provides an overview of the directions and recommendations. The recommendations are then described in detail in section III.

- It is not expected that any given funder or scientific stakeholder will aim to address all the directions below. Rather, they will focus on directions of greatest interest and relevance to them.
- It is also not expected that funders or scientific stakeholders will choose to follow the recommendations precisely. The ICDA Recommendations are intended to lay out coherent plans to help shape activities, which may help funders in defining their own programs and the scientific community in launching projects.
- We have not sought to explicitly prioritize the directions and recommendations. They are all important, and we expect that different stakeholders will have different priorities.
- As stated above, ICDA will aim to serve as a unifying forum for envisioning, stimulating, and, as appropriate, coordinating activities related to these directions and recommendations.

Directions. This framework includes 8 key directions: flagship diseases, maps, mechanisms, medicine, data, policy, global equity, and investigator-initiated innovation.

Direction 1: Flagship Diseases Projects

Flagship Disease Projects will be a crucial cross-cutting element of ICDA, serving as both driving problems and testbeds for new methods and paradigms to go from Maps to Mechanisms to Medicine for *any* common disease.

Flagship Disease Projects will help identify the most effective ways to identify (i) the variants and genes that play a causal role in disease etiology and progression; (ii) the cell types, tissues and organs in which these genes act to affect the disease; (iii) the cellular pathways and mechanisms that drive disease; (iv) potential targets for prevention, detection and therapy; and (v) the tools needed for diagnostic and therapeutic development.

Flagship Disease Projects should be launched for *at least* ten important diseases, selected based on a variety of scientific, medical and socioeconomic criteria. These diseases should also serve as testbeds for many of the cross-cutting efforts described below.

The choice of diseases will ultimately be up to scientific champions paired with funders committed to accelerating progress on the disease, working in close consultation with the scientific community.

- **Recommendation 1** describes Flagship Disease Projects.

Direction 2: Maps

Maps Projects aim to propel the genetic mapping of genetic variants that shape human diseases and traits — by improving large international biobanks and creating key genetic resources. There are five specific recommendations:

- **Recommendation 2**: Increase the diversity of biobanks focused on genetic analysis, by establishing and expanding biobanks in Africa, the Americas, and Asia.
- **Recommendation 3**: Increase the overall size and utility of biobanks, by developing more biobanks in the context of medical systems (especially with deep genetic and clinical information with electronic medical records) and enabling recontact of participants for following up clinical research and clinical trials.
- **Recommendation 4**: Propel federated genetic analysis across global biobanks, by supporting ongoing efforts, including a recently-launched network.
- **Recommendation 5**: Assemble a comprehensive genetic-variant resource inclusive of geographic and genetic diversity, including representation from Europe, East Asia, South Asia, Central Asia, West Africa, East Africa, Southern Africa, the Americas, and Oceania.
- **Recommendation 6**: Create a comprehensive, curated resource of GWAS summary statistics, providing the full information about results required for modern analyses.

Direction 3: Mechanisms

Mechanisms Projects are targeted at identifying the direct molecular and cellular functional consequences of genetic variants, enabling the identification of causal genes and pathways for disease biology. There are three specific recommendations:

- **Recommendation 7**: Create a Human Gene Regulation Map, to comprehensively connect noncoding cis-regulatory elements to their target genes across all cell types.
- **Recommendation 8**: Create a Comprehensive Catalog of Cellular Programs, consisting of coordinated changes in gene expression across cell types and states and an understanding of the regulatory elements that control them.
- **Recommendation 9**: Develop systematic characterization of all common protein-coding variants, including disease risks and cellular effects.

Direction 4: Medicine

Medicine Projects focus on developing functional assays to unravel disease-related processes, identifying biomarkers to support both clinical translation and genomically-informed clinical trials. While systematic, disease-agnostic approaches provide a very powerful foundation, disease-specific efforts will be essential to propel therapeutic development and clinical care. There are five specific recommendations:

- **Recommendation 10**: Develop functional assays for systematic dissection of disease-related processes, to implicate genes and variants in cellular processes.
- **Recommendation 11**: Create tools to propel therapeutic development for high-priority targets, to prioritize and validate mechanisms as therapeutic targets.
- **Recommendation 12**: Develop new methods for discovering biomarkers of disease, to provide mechanistic insights and support clinical translation.

- **Recommendation 13**: Accelerate genomically-informed clinical trials, to increase the success of trials while decreasing patient numbers, enrollment time and adverse events.
- **Recommendation 14**: Establish best practices for the responsible use of polygenic scores in medicine, including the creation, assessment, certification and clinical use of Polygenic Risk Scores.

Direction 5: Data

The diverse types of data and results generated across the different directions highlight the need for a federated ecosystem of software resources. Data projects will facilitate the storage, curation, access, analysis, and dissemination of data and results for use by the global community.

- **Recommendation 15**: Create a common open-source Data Platform — a shared software infrastructure that allows diverse users to create distinct Data Repositories providing data storage and data analysis, while enabling federated analyses in compliance with national regulations, IRB approvals, and data-use permissions.
- **Recommendation 16**: Create Genotype x Phenotype portals to readily access information about the relationship between genes and phenotypes, including genetic data, functional assay data, and clinical information.
- **Recommendation 17**: Create curated, freely available gold-standard validation datasets for the robust and unbiased testing, benchmarking, and validating of novel computational methods to infer causality at all steps in the process from maps to mechanisms to medicine.

Direction 6: Policy

In order to responsibly drive progress toward the M2M2M goals, it is important to have well-crafted policies to promote data sharing, protect participants, comply with national regulations, and respect cultural perspectives.

- **Recommendation 18**: Establish a Data Compliance Task Force to review policy needs and develop concrete proposals, and work with national authorities and funders to refine and implement these proposals.

Direction 7: Global Equity

The mission of understanding and treatment of common diseases must be a global activity — involving both research participants and scientific researchers around the globe.

- **Recommendation 19**: Work with funders and other ICDA stakeholders to secure equitable access to funding, training programs, and advanced technologies needed to tackle the M2M2M Challenge for researchers in lower- and middle-income countries.

Direction 8: Innovation and Training

While larger-scale efforts will be important for addressing the M2M2M Challenge, there is a tremendous need for many individually initiated efforts aimed at:

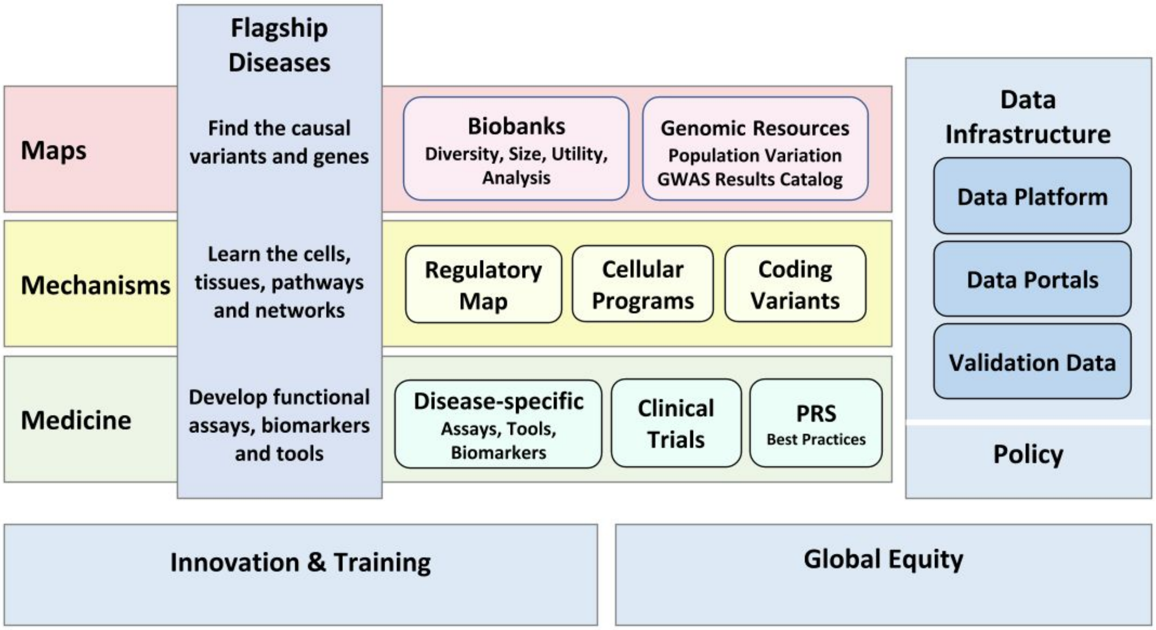
- **Recommendation 20**: Development of new analysis methods.
- **Recommendation 21:** Development of new technologies.

- **Recommendation 22:** Development of new biological directions.
- **Recommendation 23:** Training new investigators.

We note that the ICDA directions and recommendations are focused on human biology and medicine. However, we emphasize that animal models should play a key role in some of the recommendations.

Role of Ethics. Finally, we have chosen not to make a separate recommendation about "Ethics" because we don't view ethics as separable. Rather, ethical considerations must be embedded throughout each ICDA direction and recommendation.

The diagram below provides a schematic overview of the **overall framework**.



III. Detailed Description of Recommendations

Direction 1: Flagship Disease Projects

Flagship Disease Projects will be a crucial, cross-cutting element of ICDA, serving as both driving problems and testbeds for new methods and paradigms to go from Maps to Mechanisms to Medicine for *any* common disease.

While Flagship Disease Projects might logically be described *after* the recommendations concerning Maps, Mechanisms, and Medicine, we have focused on them *first* to emphasize our view that the utility of new approaches will be best measured by their ability to advance progress on the understanding and treatment of diseases.

Flagship Disease Projects will help identify the most effective ways to identify (i) the variants and genes that play a *causal* role in disease etiology and progression; (ii) the cell types, tissues and organs in which these genes act to affect the disease; (iii) the cellular pathways and mechanisms that drive disease; (iv) potential targets for prevention, detection and therapy; and (v) the tools needed for diagnostic and therapeutic development.

Flagship Disease Projects should be launched for ***at least*** ten important diseases, selected based on a variety of scientific and medical criteria. These diseases should also serve as preferred testbeds for many of the horizontal efforts described below.

The choice of diseases will ultimately be up to scientific champions paired with funders committed to accelerating progress on the disease, working in close consultation with the scientific community.

Understanding the genetic basis of a common disease means having a comprehensive picture of the genes that play a causal role in etiology and progression of the disease; the cells, tissues and organs in which the genes act to affect the disease; the cellular pathways and mechanisms that drive disease; and potential targets for prevention, detection, and therapy.

While there has been progress on many specific diseases, we have had no systematic way to obtain a comprehensive picture of a common disease.

With recent advances across many fields (including genomics, cell biology, genome editing, data science, and biobanking), the longstanding goal of having a systematic paradigm for understanding the genetic basis of common diseases is now potentially within reach.

But, tremendous work will be needed to achieve this goal.

We have powerful methods to map genetic variants underlying common diseases, but we have not yet applied them to reach a comprehensive picture for <u>any</u> disease. With respect to rare variants with large effects, which are so valuable for understanding diseases, we have not yet reached the necessary sample sizes. (Given the frequency of rare variants in human genes, this will require sequencing in the range of 250,000 cases; see below.) With respect to more common variants with more modest effects, which can together reveal biological pathways, we do not yet have genetic saturation. (Operationally, we define genetic saturation in terms of functional convergence — that is, the vast majority of disease-associated genetic variants falling into an identified set of biological pathways.)

We still lack general methods for inferring mechanisms—that is, how disease-associated variants influence disease through specific genes, cell types, cellular pathways, and organismal physiology. However, there has been steady progress in the past few years, with new general approaches based on newly available data being pioneered and tested.

We similarly lack general methods for deriving therapeutic hypotheses. However, we have an increasing number of examples in which genetic understanding is driving medical progress.

The time is right to develop general paradigms applicable to *any* common disease. The best way to do so is to undertake Flagship Disease Projects to drive a collection of common diseases toward a comprehensive picture of the nature and identity of the causal variants, genes, pathways, and mechanisms. As described in Recommendation 1 below, it is important to study a range of diseases. We propose a focus on at least 10 diseases, to be chosen by funders and scientists based on interests.

These Flagship Disease Projects will serve as high-priority focus areas for the approaches discussed subsequently — including with respect to data gathering, and in comparing new methods with respect to their effectiveness and accuracy in solving the *M2M2M Challenge*.

Recommendation and Implementation. The ICDA Organizing Committee should take direct responsibility for ICDA's activities in connection with Recommendation 1.

Recommendation 1: Launch Flagship Disease Projects, for at least 10 common diseases

Cross-cutting flagships projects should be launched for at least 10 common diseases — with the aims of:

- validating a clear roadmap that can ultimately be used for any disease to move from Maps to Mechanisms to Medicine, including by evaluating the effectiveness and reliability of different methodologies;
- saturating gene discovery with respect to rare protein-coding variants, as well as common variation, to assess the role of each gene;
- saturating the discovery of important disease-associated pathways; and
- developing assays, model systems, and biomarkers to support therapeutic development for the disease.

Flagship diseases should be chosen to cover a wide range of different medical areas and genetic architectures. Cases should:

1. have rich medical information about the disease, including, where possible, information about disease progression, treatment responses, adverse reactions, environmental exposures, phenotypic variation, and comorbidities;

2. represent a wide range of ancestries (including currently underrepresented ancestries) to learn from population differences and to ensure that results serve patients from many populations;
3. represent both sexes (with the exception of sex-limited conditions); and
4. be subjected to whole-genome sequencing.

Cases should be analysed together with ancestry-matched shared controls, to be used across many diseases.

For typical common diseases, projects should analyze at least ~125,000 cases with an ultimate goal of analyzing ~250,000 cases.

These Flagship Disease Projects will also serve as high-priority focus areas for the approaches discussed below — including with respect to data collection and comparing new methods with respect to their effectiveness and accuracy in solving the *M2M2M Challenge*.

To advance this goal,

1. **ICDA, in collaboration with potential funders and other stakeholders, should engage the scientific community and potential funders concerning:**
    a. **which diseases would be most appropriate for Flagship Disease Projects, and**
    b. **how such projects should best be designed and executed.**
2. **Funders should support well-chosen Flagship Disease Projects to drive progress on specific diseases of interest to them and to propel development of a general paradigm applicable to any disease. (Each Flagship Disease Project might be supported by a single funder or by a group of funders.)**

The initial Flagship Disease Projects are intended to serve as exemplars and testbeds for how to go from Maps to Mechanisms to Medicine. It is important to have *at least* 10 diseases — spanning a range of medical areas, organ systems, genetic architectures and technical challenges — to ensure we meet and overcome the key challenges and develop a paradigm to tackle *any* disease.

Importantly, this is not intended to *exclude* diseases — more than 10 diseases can be included if there is sufficient interest and funding.

Choice of diseases. The diseases should be chosen based on a range of considerations relevant to maps, mechanisms, and medicines, including:

- **medical importance, unmet need, burden on global health, and relevant comorbidities**;
- **existing progress in genetic discovery**, through active disease-specific consortia;
- **availability of many samples from disease cases, including from diverse populations and both sexes (except for sex-limited diseases);**
- **accessibility of relevant cell types**;
- **current knowledge of biological processes and intermediate phenotypes;**
- **availability of cellular models**;
- **pharmaceutical interest, clinical trial infrastructure, and regulatory landscape**; and

- **commitment to public deposit of genotype and phenotype data.**

Importantly, the choice of diseases will be strongly influenced by stakeholders and funders who want to drive progress on particular diseases.

It is not intended that diseases meet all of these criteria. Rather, the diseases will ideally meet only *some* of the criteria — so that we learn how to overcome challenges.

The set of diseases should ideally cover a range of areas. Examples could include autoimmune and inflammatory disease, cardiometabolic disease, cancer susceptibility, neurodegenerative and neuropsychiatric disease, respiratory disease, host genetics of infectious disease, and reproductive and women's health conditions.

It will be important to work on ways to harmonize phenotype data from routine healthcare data.

Number of cases for genetic analysis. Flagship Disease Projects should aim for approaching genetic saturation of genes that can be identified by both common and rare variants. For typical diseases, the Maps component will require genetic analysis of ~250,000 cases, which should represent diverse ancestry around the globe. As an intermediate goal, projects might aim for 125,000 cases.

Large numbers of cases are important for several reasons. They will be essential for identifying genetic influences not only for susceptibility, but also for disease onset, progression, course, and treatment response. Large samples are also required for analyzing disease subtypes based on disease-related phenotypes, for studying important differences across ancestries, and for developing polygenic risk scores appropriate for different populations.

They are also essential for ensuring that there is sufficient power to discover rare protein-coding variants with large effects in the substantial majority of genes. Rare alleles are tremendously valuable for biological and medical purposes — especially loss-of-function (LoF or knockout) alleles, which reveal crucial information about gene function, direction of effect, and allelic series. The recommended sample size (~250,000 cases) is based on the frequency of LoF mutations in human genes: Roughly two-thirds of human genes have LoF frequency ≥ 3.2 per 100,000 chromosomes — corresponding to 16 occurrences per 250,000 people, which should allow detection of several-fold increased risk in cases. (Some complex diseases, such as autism, are associated by LoF variants with very large effects in mutation-intolerant genes. In such cases, many fewer cases may suffice.)

Assembling and analyzing large case-control collections. Assembling the number of cases required for Flagship Disease Projects to identify critical genes harboring rare variants of larger effect will <u>continue to require focused case-control collections</u>. At some point in the future, population biobanks may become so massive as to obviate the need for disease-focused collections in most diseases, but this remains many years out except for extremely common diseases.

The best prospect for efficient identification of cases is likely to be through medical systems with excellent health records.

The cost for assembling and performing genetic analysis of case collections will depend on the efficiency of the infrastructure for identifying cases, collecting samples, processing samples (extract plasma, white-blood cells and DNA), obtaining clinical information, and performing whole-genome sequencing (WGS). The cost of sample collection and processing is modest (the Estonian Biobank and FinnGen Projects cite costs of $40-80 per sample, respectively) and the cost of WGS is expected to continue to fall to the range of ~$150 in the next few years. The costs of identifying cases, arranging visits to collect samples, gathering medical records and analyzing data are harder to estimate.

Finally, Flagship Disease Projects should be designed to use shared controls, thereby amortizing the cost for controls over many projects.

COVID-19 Host Genetics Project

While these recommendations were being finalized in the midst of the emerging COVID-19 pandemic, it became clear that understanding the genetic contributions to the substantial variability observed in response to SARS-CoV-2 infection might represent an urgent Flagship Disease Project. The recently launched COVID-19 Host Genetics Initiative (https://www.covid19hg.org/) has already brought together over 1,000 scientists and collaborators from the human genetics community to understand the genetic influences on susceptibility, course, and outcomes, based on robust study designs that account for potential sampling biases in the data. Consequently, ICDA is facilitating this effort by bringing together the community, organizing meetings, and disseminating protocols for data collection and analysis to help enable a truly international collaborative endeavor.

Direction 2: Maps — Biobanks

The past decade has seen the emergence of population biobanks as a key resource for advancing human genetic discovery. By biobank, we mean a large collection of samples and accompanying information assembled from individuals chosen without a focus on identifying cases for a particular disease. Biobanks may consist of individuals chosen from a particular population or medical system.

Pioneering work by deCODE, UK Biobank, BioBank Japan, FinnGen, several major medical centers (including Vanderbilt, Geisinger, and Mt Sinai Hospital), and others have illustrated the power of biobanks — combining large-scale recruitment, deep phenotypic information from population registries and electronic health records, and genetic information. In addition, the US has recently launched the *All of Us Project*.

Biobanks can enable many types of human genetic studies—including genetic mapping of quantitative traits, very common phenotypes, and highly prevalent diseases (for example, diabetes, present in ~6-10% of adults in Western populations), as well as clinical studies of individuals with specific genotypes.

Biobanks, however, are underpowered for genetic analysis of most diseases. A random sample of the population would be expected to contain only a limited number of cases of any particular disease, and current biobanks tend to be somewhat skewed toward healthier individuals. For example, the UK Biobank's 500,000 participants would be expected to include only 5,000 cases for a disease with 1% prevalence — far below the tens of thousands of cases needed for initial analysis, let alone the 250,000 cases ultimately desired for deep analysis (see

Recommendation 1). Moreover, some diseases are greatly underrepresented compared to expectation. For example, schizophrenia has ~1% prevalence, but fewer than 400 cases in the UK Biobank. Neurodegenerative and ophthalmologic diseases of aging are similarly underrepresented.

As the aggregate size of global biobanks grows, though, it may eventually become possible to genetically dissect *most* common diseases or traits based on biobank samples — without the need for disease-specific case-control collections. Still, it will likely still be important to maintain focused case collections for some diseases — because they may offer more detailed disease-specific phenotypes and will be necessary for diseases that are seriously underrepresented in biobanks. Moreover, with access to detailed, longitudinal, and harmonized medical records, it will be important to study not only disease incidence, but also disease progression, and response to treatments. This will involve addressing the additional challenges that are faced in extracting such information from EHRs and biobanks in a way that ensures harmonization and enables integration across data sets.

Desired Scale. What would it take to achieve this goal?

Global biobanks with a total of 25 million participants would provide 250,000 cases for genetic diseases with frequency at frequency ≥1%, if diseases were represented at their expected frequencies. In practice, a better target might be 50 million or more participants, because diseases are often underrepresented.

Such a goal is realistic. Announced targets for existing biobanks already total to roughly 8 million participants (including a large UK effort). Moreover, the coupling of biobanks to hospitals and medical systems could offer the opportunity to greatly expand the number of participants.

Expanding diversity in underrepresented regions and populations. A major issue that must be solved, however, is the heavy bias of existing large-scale biobanks towards individuals of European and East Asian ancestries. The NIH *All of Us* Project is thoughtfully designed to have greater diversity of ancestry, but its scope is currently limited to only 1 million total participants chosen from the United States. In short, much of the world's population will not be served by current efforts.

Diversity of ancestry is important for both scientific studies and clinical application: (i) the different demographic history of populations can lead to the identification of novel hits (such as the common variant in *SLC16A11* associated with type 2 diabetes, discovered in studies of the Mexican population), (ii) the different patterns of linkage disequilibrium in different populations will substantially improve genetic fine mapping, and (iii) polygenic risk scores perform best when they are generated from studies of individuals of the same population. Additional efforts should also be directed to 'special populations' — for example, populations with high rates of homozygosity due to consanguineous marriage practices, or populations with extreme founder effects. These populations are expected to be enriched for highly informative genotypes (such as heterozygosity or homozygosity of alleles with strong effects on gene function), such that even smaller biobanks from these populations may have substantial utility.

Dedicated efforts are needed to create and expand biobanks designed for genetic analysis in various regions and groups:

(i) Africa. Africa has distinctive advantages for genetic studies, because population histories in Africa have resulted in greater genetic diversity, which will reveal many new disease-associated alleles and genes, and shorter-range linkage disequilibrium, which will be invaluable for improving fine mapping of the tens of thousands of disease-associated variants that have already been discovered.

There are currently active discussions concerning creating and expanding biobanks in Africa, involving both regional scientists and organizations (e.g., the African Academy of Sciences and the African Society of Human Genetics) and global funders (e.g., Wellcome and the US National Institutes of Health (NIH)) interested in scoping possible activities.

A good initial goal might be to assemble a network of biobanks with a total of ~1 million participants from a number of geographically distinct regions. As a first step, efforts should be undertaken to develop a potential plan by the end of 2020.

(ii) Americas. In these recommendations, “Americas” will be used to refer to the underrepresented countries and populations living in South, Central and North America — effectively, all countries other than the US and Canada, as well as systematically undersampled groups in the US and Canada.

The Americas are valuable for genetic studies, because the population histories of these regions and individuals include many population bottlenecks and complex admixture. The bottlenecks and founder effects will simplify the allelic spectrum, while the admixture provides opportunities for exploring epistasis.

Several international efforts have sought to study genomic diversity in the Americas, with the aim of understanding the demographic history and diversity in the continent. There have also been local efforts, such as national genome projects in Mexico, Peru, Brazil, Argentina, and Chile, creating valuable DNA collections and genome datasets. Recently, national biobanks have been created in some countries of the regions, such as in Mexico, Brazil and Argentina and there have been efforts to connect these and smaller biobanks in a Network of Biobanks from Latin American and the Caribbean (referred to as REBLAC, its acronym in Spanish). With few resources, these efforts have created collections of small-to-medium size but with good data quality.

Given strong interest and leadership by regional scientists, it is now the time to develop a coordinated plan and to identify potential funding. A good initial goal might be to assemble a network of biobanks with a total of ~1 million participants from a number of geographically distinct regions. As a first step, efforts should be undertaken to develop a potential plan by the end of 2020.

(iii) Asia. While there are several major biobanking efforts involving Japan and China, there remain major gaps. Biobanks in South Asia are currently quite limited; the Indian government and scientific community might take a lead in such efforts. Biobanks in China could also be dramatically expanded. In particular, there is considerable interest in Hong Kong in creating biobanks. Further elucidation of human genetics in the Arab and central Asia populations is also warranted.

In addition, effort should be considered in Oceania, particularly given the multiple population bottlenecks and migration events.

Planning efforts should be undertaken, drawing on the extensive scientific communities and resources in these countries.

Expanding diversity in existing major biobanks. Even in regions where large biobanks have been developed, these biobanks often do not represent the ancestry diversity of the population. It is important to address this scientific gap, which has been identified as a priority area for NIH, Wellcome, and others. As noted above, ancestry diversity among participants in many biobanks can reveal clues to disease mechanisms, is necessary to ensure appropriately calibrated estimates of genetic risks for all individuals, and can facilitate the analysis of gene x environment interaction.

At the same time, due to historical lack of equality in research and healthcare, diverse populations often include vulnerable individuals. Considerable thoughtfulness is needed with respect to ethical and cultural considerations in recruitment and to ensuring benefit for all.

Expanding size and utility, by creating biobanks coupled to medical systems. While a number of important biobanks (notably, the UK Biobank) have been standalone activities, the most cost-effective approach to create biobanks with diverse participants and high-quality clinical information may ultimately be to couple them in the context of hospitals and medical systems with electronic medical records. Such biobanks should endeavor to capture socioeconomic and environmental data, which may play key roles in clinical outcomes. It will be important to explore and expand biobank models based on medical systems.

Enabling recall of participants by genotype, where possible. The ability to recall some participants based on their genotypes or phenotypes would be valuable, because it would enable in-depth follow-up studies for research and clinical trials. Where possible (especially in the context of medical systems), it would be desirable to ask patients for their consent to be recontacted for potential studies. In addition, it would be desirable to seek consent for return of results under clear policies. This will be particularly important in studies involving individuals with rare genotypes (such as those who carry homozygous rare loss of function alleles).

Genetic analysis network across global biobanks. Realizing the power of biobanks will require that appropriate ways are developed to allow joint analyses across many disparate biobanks in a manner that delivers robust and definitive analyses to the entire research community.

This will require investment in phenotypic harmonization across both new and existing biobanks — performed in both traditional ways but also by using opportunities provided by progress in human genetics itself. Specifically, by using established risk loci and polygenic scores as instruments, the consistency of phenotype definition can be assessed across biobanks in an orthogonal way that speaks directly to the utility of joint genetic analyses.

The potential of a biobank network will be particularly impactful in yet-to-be-explored areas, such as response to therapy and disease progression, where lack of real-world data in traditional study designs and the challenge of starting with smaller populations with specific diagnoses, age categories and treatments have been limiting.

To realize the full potential of cross-biobank analyses, a robust global biobank analysis network will be essential.

A **Global Biobank Network** was initiated in 2019, with the aim of creating a framework in which phenotype harmonization and meta-analysis can be performed in a facile and collaborative framework, with rapid and open sharing of results. Early pilot studies, shown at American Society of Human Genetics (ASHG) in October 2019, demonstrate the potential of such a network. For example, joint analysis of 82,000 asthma cases across 9 biobanks revealed 82 genome-wide significant findings. It will be important to nurture and support these nascent efforts.

Coordination with IHCC. The International Hundred-K+ Cohorts Consortium (IHCC) was launched in 2018 to create a global network for translational research using large cohorts. IHCC aims to bring together large cohorts to encourage data sharing, improve efficiencies, and maximize benefits in addressing scientific questions.

IHCC has a broader scope than ICDA, including many cohorts focused on epidemiology rather than genetic analysis of common disease. It will be important for ICDA to coordinate closely with IHCC. In particular, ICDA should focus on issues related to genetic studies with large biobanks.

Recommendations and Implementation. Concerning biobanks, we make three recommendations (Recommendations 2, 3, and 4).

ICDA's activities in connection with these recommendations should be coordinated by a **Biobanks Task Force** within the ICDA Maps Working Group.

Recommendation 2: Increase the Population Diversity of Biobanks focused on Genetic Analysis

The population diversity of biobanks designed for genetic analysis should be substantially increased.

An early goal should be to support nascent and existing efforts to develop biobanks in regions that have been seriously underrepresented. Initial planning should begin with active efforts in Africa and the Americas (see definition above), and should embrace additional efforts in South Asia and East Asia, as they are developed.

Additionally, efforts should be made to increase the diversity of participants in biobanks *within* developed countries, to ensure that all individuals maximally benefit from research and to reduce healthcare disparities.

To advance this goal,

1. **Funders should facilitate, support and set clear targets for:**
   a. **expansion and creation of biobanks in underrepresented regions; and**
   b. **greater recruitment of underrepresented populations into existing major biobanks.**
2. **ICDA should help to support and expand existing biobanks and build new biobanks, by working with the regional scientific communities, global experts, potential funders, IHCC, and others to:**
   a. **convene meetings to map the scope of projects, including how to carry them out most effectively and efficiently in the regional context;**

b. define the necessary resources (logistical, budgetary, analytical, computational);
c. ensure that these efforts can maximally benefit from existing knowledge, protocols and infrastructure; and
d. develop concrete plans and milestones by the end of 2020.

Planning should begin with nascent efforts in Africa and the Americas, and should embrace additional efforts in South Asia and East Asia, where they have strong local support.

Efforts to create or expand biobanks in underrepresented regions should be led by regional scientists assisted by external partners (rather than vice versa, e.g., 'safari science'). They should include clear commitments to local leadership, to advancing equity, and to capacity building, including education, training, and infrastructural investments.

Efforts should bring together existing research groups with common goals and recruitment capacities, and they should draw on existing strengths. For example, projects might be built in the context of hospitals and healthcare systems rather than as standalone biobanking efforts — which might increase the representation of diverse populations, advance the development of electronic medical records, and facilitate the ability to recontact patients for in-depth follow-on studies.

Recommendation 3: Increase the Size and Utility of Biobanks focused on Genetic Analysis

An ultimate goal should be a federation of biobanks across the globe at sufficient scale to enable deep analysis of the vast majority of common diseases and disease-related phenotypes. This will likely require a total of at least 50 million participants.

To be maximally useful, these biobanks will need to have rich clinical and genetic information — including clinical trajectories, responses, and outcomes for patients with the same disease. The most effective way to assemble such biobanks will be in the context of hospitals and healthcare systems with excellent electronic health records containing rich longitudinal, multidimensional information, combined with appropriate methods to extract relevant phenotype and risk factor information from the health records.

Information in electronic health records will make it possible to learn about distinct aspects of disease processes, including those with the greatest relevance to therapy. (Stakeholders, including pharmaceutical companies, should work together to maximize the ability to combine outcome data from different sources — including through harmonization of outcomes measures and genetic cross-validation of phenotypic standards.)

Wherever feasible, biobanks should collect data on environmental and social determinants of health, and enable participants to consent to recontact for future study.

There should be particular attention to expand and coordinate the identification of individuals with rare loss-of-function alleles (especially those present in the homozygous state) from diverse ancestral backgrounds, and to develop a coordinated infrastructure to support detailed clinical investigation in such individuals and their families.

To advance this goal,

1. **ICDA should, in collaboration with potential funders and stakeholders, work to identify opportunities, barriers, and pilot projects for developing effective biobanks in the context of hospitals and healthcare systems around the world — especially, where possible, those with rich medical records and the ability to enable patients to consent to recontact for future studies. In developing such biobanks, it will be important to develop effective ways to engage patients, align institutional interests, design policies, use existing infrastructure, and create required new infrastructure.**
2. **Funders should support pilot projects to design and expand effective biobanks in the context of hospitals and healthcare systems.**

Recommendation 4: Support Genetic Analysis Across Global Biobanks

To fulfill the potential of biobanks for understanding diseases, it will be necessary to be able to perform federated genetic analysis of diseases and traits across global biobanks.

This federation of biobanks should work to:

- align data processing, quality control, and imputation references; and
- integrate clinical and other phenotypes, such as environmental exposures, including developing novel and rigorous ways to ensure that cross-biobank analyses are as accurate as possible — for example, by using genetics to cross-validate phenotypic definitions.
- explore phenotypes that have been difficult to explore in traditional study designs and limited sample sizes, such as disease subcategories, disease progression and trajectories, treatment response, sex differences, and age categories.

To advance this goal,

1. **Funders should support efforts to create a genetic analysis network to facilitate and accelerate federated analysis across global biobanks.**
2. **ICDA should support a genetic analysis network across global biobanks by helping to facilitate and support meetings of the nascent Global Biobank Network.**

Direction 2: Maps — Genetic Resources

Certain resources play a crucial role in human genetics research, including reference catalogs of (i) the genetic variation in populations, and (ii) the results from genetic mapping studies.

Comprehensive genomic variations resources. Freely available human genetic reference datasets are an essential foundation for genetic mapping of human complex traits.

The International HapMap Consortium, the 1000 Genomes Project, gnomAD, and Haplotype Reference Consortium are among the most widely used resources in the human genetics community.

These comprehensive genomic variation resources are required for several critical purposes:

- They are used to create allele frequency servers (such as gnomAD) to help medical geneticists around the world to evaluate patients suspected of having a Mendelian disease — enabling them to focus on likely causal mutations by filtering out the vast majority of variants that are sufficiently common in one or more major populations.
- They are used to create imputation servers and linkage-disequilibrium reference panels to facilitate disease-mapping studies — by using information about haplotype structure to infer genotypes at many millions of common variants based on genotypes at a much smaller set of variants obtained from genotyping arrays.
- They reveal the patterns of natural selection in each gene, including the frequency of LoF alleles. Notably, these patterns may be population-specific.

The value of these resources increase with larger and more diverse samples. Sample sizes of 100,000 for every major ancestry group will enable genotype imputation down to a minor allele frequency of ~ 0.1%, as well as provide a necessary reference for clinical variant interpretation.

Current genetic resources focus primarily on SNPs and small indels, which constitute the vast majority of human genetic variants. More complex variation can also play important roles in disease, but is currently more difficult to assay. While genotyping of such variants is currently too expensive for large-scale implementation, efforts should be made to develop ways to enable genotyping of more complex variation without significantly increasing costs.

Genome-wide association summary statistics resources. Resources that assemble the results from the thousands of genetic-mapping studies performed to date play a critical role in human genetics.

In particular, well-curated GWAS summary statistics are essential because they help investigators to:

- readily identify and assess genes and variants to target for experimental investigation of a complex disease;
- compare association results across phenotypes;
- enable Mendelian Randomization and other types of causal inference analyses;
- define allelic series needed to describe, for any gene, the relationship among genetic variants (coding and non-coding), their impact on molecular and cellular function, and their phenotypic consequences (both for desired outcomes and adverse effects) — including defining the genetic “dose-response” curve; and
- create polygenic risk scores.

Early efforts to catalog the results of GWAS by NCBI and EBI have been invaluable.

However, the current catalogs have serious limitations. Some examples: The catalog only contains the top hits for many GWAS, rather than full genome-wide association results. For ongoing consortium efforts, such as GiANT and the Psychiatric Genomics Consortium, the catalog often lists results from multiple overlapping meta-analyses — leading to confusion about which version is the most accurate and up-to-date. Clear standards for different GWAS outputs, such as fine-mapping, gene prioritization, and SNP weights for PRS, have not yet been developed or disseminated. Efforts are underway, including at EBI and in the Netherlands, to address some of these issues, but much remains to be done.

Recommendations and Implementation. Concerning Genetic Resources, we make two recommendations (Recommendations 5 and 6).

ICDA's activities in connection with these recommendations should be coordinated by a **Genetic Resources Task Force** within the ICDA Maps Working Group.

Recommendation 5: Create a Comprehensive Genomic Variation Resource

Comprehensive genomic variation resources — consisting of organized and readily accessible genetic data based on deep whole-genome sequence data from many individuals across many population groups — have been an essential foundation for human disease genetics (*see above*).

To support these crucial purposes, we need to expand existing resources to create a comprehensive genomic-variation resource based on whole-genome sequence for at least one million individuals representing an expansive range of geographies in order to capture as much ancestral diversity as possible (for example, ~100,000 from Europe, East Asia, South Asia, Central Asia, West Africa, East Africa, Southern Africa, North America, Central America, South America, and Oceania).

Existing resources fall far short of the necessary coverage for many major ancestry groups. Currently, the target size of 100,000 has been achieved only for European ancestry. For the other major ancestry groups above, the available data is much smaller—often only in the thousands—due both to limitations in data collection and in data sharing. Addressing these gaps is essential for addressing ancestry-based disparities in genetic diagnostics.

To the greatest extent possible, samples should be chosen from across each region.

Deeper sampling may be ultimately appropriate for populations with higher genetic diversity, such as Africa, and with many population bottlenecks, such as India. These choices should be guided by the initial results.

Where possible, samples should come from existing projects. However, new samples will likely need to be generated to provide appropriate population coverage in some cases.

To advance these goals,

1. **ICDA, in collaboration with regional stakeholders and funders, should work to:**
    a. **design and support efforts to generate comprehensive genomic variation resources from underrepresented ancestry groups;**
    b. **ensure that, to the greatest extent possible, individual-level genomic information (without phenotypic information) is consented for broad sharing and integration, to maximize the utility of the data for the scientific community;**

   c. **promote the continued advancement of allele frequency servers for medical genetics, including engaging international stakeholders to ensure that the servers meet the needs of many countries and have a sustainable support mechanism;**
   d. **promote the continued advancement of imputation servers; and**
   e. **facilitate the ability to use individual-level data as genetically matched controls for disease studies.**
2. **Funders should provide support to ensure the creation and accessibility of the foundational genetic resources needed for comprehensive studies of human disease.**

Recommendation 6: Create a curated GWAS Summary Statistics Resource

A comprehensive, readily accessible resource of all previous GWAS results is crucial for disease studies. (Existing GWAS resources have been valuable in allowing investigators to search the results of individual papers, but they do not integrate and reconcile the results to provide the information needed for follow-up genetic, functional and clinical studies.)

A comprehensive, public GWAS repository and server should include:

- complete genome-wide summary statistics for every association study;
- the current best-powered meta-analyses for each disease or trait, generated collaboratively with disease consortia (superseding earlier meta-analyses);
- credible sets for each reliable association in the meta-analysis;
- polygenic risk estimators for diseases and traits (with weights and relevant metadata) that have been used or proposed for research or clinical purposes; and
- Application program interfaces (APIs) for easy access, filtering, analysis display of results, and bulk download of summary statistics.

Discussions are already underway concerning the creation of such a resource, including at the European Bioinformatics Institute (EBI).

To advance this goal,

1. **ICDA should:**
   a. **work with key stakeholders and ongoing efforts (such as at EBI), to determine how a comprehensive GWAS Summary Statistics Resource should best be designed and executed; and**
   b. **work with disease consortia, biobanks, funders, and journals to design and implement policies to develop clear standards for summary statistics to be deposited and ensure that all relevant projects deposit their summary statistics (see also Recommendation 18).**
2. **Funders should support the creation and maintenance of a comprehensive GWAS Summary Statistics Resource, as described above.**

The complete genome-wide summary statistics should include variation types, risk allele, effect size, direction of effect, sample size, and metadata about the generation of the summary statistics (including genome reference build, imputation reference used, linkage-disequilibrium information, association testing model, and phenotype definition for any associated phenotypes linked with electronic medical record codes).

Direction 3: Mechanisms

For most of the more than 70,000 loci that have been associated with diseases and traits, the mechanisms through which they affect disease risk remain unknown. This is a crucial part of the *M2M2M Challenge*.

The term “mechanisms” is regularly used to refer to several different aspects of disease biology, including: (1) the *direct* effect of a variant on its immediate target gene (often called *cis* effects); (2) the *downstream* effect of a variant on gene regulatory programs (often called *trans* effects), as well as other biochemical programs, within a cell; and (3) the physiological consequences of a variant for the function of cells and tissues, and for the health and disease of the organism. The first two meanings are the focus on this section (“Cellular Mechanisms”), while the third meaning is considered in a subsequent section (“Medicine”).

For the ~90% of these loci that are not driven by coding variants, a major roadblock is our relatively poor knowledge of the regulatory code of the human genome. Projects such as ENCODE and GTEx have made major contributions, but they have been limited by sample size, representation, and technologies — which lacked cellular resolution and experimental perturbation.

The ~10% of loci that involve coding variants in known genes are particularly valuable for studying physiology and medicine. Yet, we have lacked systematic ways to characterize the downstream cellular programs that are affected. Moreover, functional-genetics studies to date have provided poor representation of non-European ancestries.

We need to dramatically improve our ability to connect genetic variants to cellular mechanisms. Fortunately, powerful new approaches offer the opportunity to create comprehensive resources that would dramatically accelerate progress. These approaches include observational analysis of human biospecimens at single-cell resolution that might eventually be amenable to measuring *all* cell types and states and experimental perturbations, typically involving CRISPR-based approaches, that can be applied in more limited setting but that allow rigorous testing and can support the development of broadly applicable predictive models.

To accelerate the progress from Maps to Mechanisms, we believe it is time to begin three important activities:

- Create a **Human Gene Regulation Map**, to comprehensively connect *cis*-regulatory elements to their target genes in all cell types, by combining single-eQTL analysis, CRISPR perturbations, and computational modeling.
- Create a **Comprehensive Catalog of Cellular Programs**, to identify coordinated changes in gene expression caused by the *trans*-activity of sets of transcription factors

across cell types and states — including determining the transcription factors, regulatory elements, and target genes — by combining single-cell molecular analysis with natural genetic variation and experimental perturbations; and

- Undertake **Systematic Characterization of Protein-Coding Variants,** to identify disease risks associated with common variants (all variants with frequency ≥ 0.1%) and cellular effects associated with variants (disease-associated common and rare variants, as well as knockout alleles).

These activities are intended to use high-throughput methods to create *comprehensive* resources that would ultimately be applicable across the entire genome and across all cell types, in both sexes. We note that the activities would be complemented by lower-throughput disease-specific approaches discussed under the *Medicine* section below.

The activities are clearly ambitious, but we believe that they can all be accomplished within a 5-10 year time frame. It would be sensible to start by focusing on a subset of variants and cell types, which could be prioritized in part based on their relevance to the Flagship Disease Projects.

Relationship to other efforts. ICDA's goals in this area will require an alliance of the human genetics, single-cell, gene regulation, and epigenomics communities.

There are a number of important existing efforts — including the Human Cell Atlas, GTEx and ENCODE — with related goals, although with a somewhat different focus than ICDA. These communities are actively engaged in technology development, large-scale data generation, innovation in computational methods and modeling, and pilot projects.

ICDA's efforts, from planning to execution, should be closely coordinated with these efforts.

Recommendations and Implementation. We describe these three projects in three recommendations below (Recommendations 7, 8, and 9). ICDA's activities related to these recommendations should be coordinated by the **ICDA Mechanisms Working Group**.

Recommendation 7: Create a Human Gene Regulation Map

A collaborative project should be launched to build a comprehensive map of *cis*-regulatory elements across all cell types as a crucial resource for connecting disease-associated non-coding variants to the nearby target genes that they regulate in *cis* in each cell type

Across *all* human cell types, the project would aim to:

- Identify all cis-regulatory elements (cREs),
- Infer the target genes affected by the cREs in *any* cell type,
- Identify all cis-QTLs having effects above a threshold;
- Infer the functional architecture of cREs (e.g., binding sites for specific proteins) to allow predictive models of effects of variants on cRE function and gene regulation.

This information could be gleaned from various data (including gene expression, transcript splicing, chromatin assays, CRISPR-based perturbation of gene expression in selected cell types, and computational models created from these data).

Studies should involve samples that are diverse with respect to ancestry and involve both sexes.

To advance this goal,

1. **ICDA should work with scientists, potential funders, and existing projects — especially the Human Cell Atlas and ENCODE — to:**
    a. **determine how the project might best be designed and executed over time, including how to rigorously compare and validate different methods;**
    b. **convene ongoing scientific meetings to assess progress; and**
    c. **develop, validate and share SOPs for collection and creation of biospecimens and cell line and organoid resources from diverse populations.**
2. **Funders should support efforts to create a Human Gene Regulation Map, beginning with pilot projects and expanding as appropriate.**

Many of the relevant approaches have been demonstrated at small scale, but important issues remain about how best to deploy them to create a comprehensive resource.

Samples. For studies of natural genetic variation, samples can, in principle, be collected from any tissue via biopsy and/or autopsy (as in the GTEx Project), although initial studies will be influenced by ease of procurement. Many hundreds or thousands of samples will be needed to reliably detect and estimate disease-relevant effect sizes. For studies of experimental perturbation, it will be necessary to use feasible human cellular models (such as readily accessible primary tissues, high-fidelity cellular and organoid models, and differentiated iPSCs). Selection should consider feasible scale of data generation, relevance to disease, and feasibility of genome editing.

Assays. Assays involve single-cell RNA expression in cell types from tissue samples or CRISPRi perturbation of cellular models.

Recommendation 8: Create a Comprehensive Catalog of Cellular Programs

A collaborative project should be launched to assemble a Comprehensive Catalog of Cellular Programs — that is, coordinated *trans*-activation in gene expression driven by sets of transcription factors turned on or off in different cell types, states, and differentiation pathways. Such a catalog would be an invaluable resource in connecting the perturbation of disease-associated genes to their downstream cellular consequences.

A project might systematically discover cellular programs based on studying how gene expression and other functional readouts at the single-cell level are affected by:

- genetic variation across many individuals—including common variants, rare loss-of-function and other coding alleles, and variation in polygenic risk scores—which can, in principle, be observed across all cell types and variants in the human population; and

- experimental perturbation—such as CRISPRi-based inhibition of genes, isogenic substitution by genome editing, expression of open reading frame (ORF) libraries with alternative common coding variants, and treatment with drugs and growth factors—which can be performed in many cell types and for many variants, but likely not comprehensively.

Given limitations in human studies (e.g., concerning natural variation and experimental validation), it will be valuable to conduct parallel studies in key model organisms, including population-based model organism models when appropriate.

Human studies should involve samples that are diverse with respect to ancestry, and human and animal studies should involve both sexes.

To advance this goal,

1. **ICDA should work with scientists, potential funders, and existing projects — especially the Human Cell Atlas— to:**
    a. **determine how the project might best be designed and executed over time, including how to rigorously compare and validate different methods;**
    b. **convene ongoing scientific meetings to evaluate progress; and**
    c. **develop, validate and share SOPs for collection and creation of biospecimens and cell line and organoid resources from diverse populations and particularly informative genotypes (such as human knock-outs).**
2. **Funders should support efforts to create a Comprehensive Catalog of Cellular Programs, beginning with pilot projects and expanding as appropriate.**

Many of the relevant approaches have been demonstrated at small scale, but important issues remain about how best to deploy them to create a comprehensive resource.

Samples. Similar considerations apply as for Recommendation 7. In addition, pooled cells from many individuals (such as iPSCs, differentiated according to various protocols) may prove valuable for sensitive detection of natural genetic variation controlling cellular programs. Existing banks of iPSCs and primary cells will be valuable resources, and efforts should engage with various existing collaborations focusing on iPSCs. These resources, however, need to be increased in size and diversity.

Assays. It will be valuable to evaluate the utility of high-content, high-throughput assays for cellular phenotyping, including single-cell RNA-seq and cellular morphological imaging. Pilot projects will need to explore how best to combine such perturbations and readouts to establish cellular programs, how best to computationally define cellular programs from high-content measurements, and how best to apply these approaches to understand the functions of disease variants and genes.

Recommendation 9: Enable Systematic Characterization of Protein-Coding Variants

Among disease-associated variants, coding variants have particular value for biological understanding and clinical translation. The number of common coding variants in populations is tractable (for example, roughly 75,000 coding variants with frequency ≥0.1% across Europe), making these variants amenable for comprehensive analysis.

A collaborative project should be launched to systematically advance our understanding of how coding variants affect gene function.

Potential approaches would include:

- Assessing the effects of common coding variants by complete PheWAS analysis across all biobanks to identify variants likely to have important functional effects;
- Assessing the effects of both common and rare coding variants, by using isogenic genome editing at large-scale to create libraries of cells carrying these variants and subjecting them to a wide range of biological characterizations and functional assays;
- Performing high-throughput saturation mutagenesis of selected genes of high interest, coupled to generic or gene-specific assays;
- Testing selected coding variants in model organisms, including mouse; and
- Developing and calibrating improved computational methods for predicting coding variant effects on protein structure, protein interactions, and function.

Studies should involve samples that are diverse with respect to ancestry and involve both sexes.

To advance this goal,

1. **ICDA should work with scientists, potential funders, and existing projects to:**
   a. **determine how the project might best be designed and executed over time, including how to rigorously compare and validate different methods;**
   b. **convene ongoing scientific meetings to evaluate progress; and**
   c. **develop, validate and share SOPs for collection and creation of biospecimens and cell line and organoid resources from diverse populations.**
2. **Funders should support efforts to systematically characterize protein-coding variants, including pilot projects and larger efforts as appropriate.**

Many of the relevant approaches have been demonstrated at small scale, but important issues remain about how best to deploy them to create a comprehensive resource.

Samples. Similar considerations apply as for Recommendation 7.

Assays. Assays for the effects of coding variants could include generic readouts (e.g., assessing overall cellular gene expression with single-cell RNA-seq, including recognizing functionally null, hypomorphic, hypermorphic, and neomorphic alleles; assessing effects on the level and splicing of a gene's own transcript; and assessing protein stability by GFP fusions) and a selection of disease-specific or gene-specific assays (e.g., autophagy reporter for IBD).

Variants. The variants characterized should include all common coding variants (frequency ≥ 0.1%), disease-associated rare variants and knock-out alleles for all genes.

Direction 4: Medicine

Understanding the genetic basis of common disease is tremendously important for the development of therapies and the provision of clinical care.

- It is becoming increasingly clear that the probability of success is highest when therapies are based on a genetically-defined mechanism of action and are quantitatively matched to specific physiological processes.
- It is also becoming clear that clinical management of patient groups informed by their genotype (stratified medicine) has great potential to direct medical resources to patients who are at higher risk and would not otherwise receive optimal care.

The Cellular Mechanisms section describes systematic, disease-agnostic approaches that will have an important impact on our understanding across all diseases and traits. However, targeted, disease-specific approaches will also be essential for the development of therapies for specific diseases and the refinement of clinical care. While disease-specific approaches necessarily have lower throughput, there is no substitute for the information that they can provide.

We discuss below four important areas of focus for therapeutic development:

- Develop ***Functional assays for systematic dissection of disease-related processes*****,** to implicate genes and variants in cellular processes relevant to particular diseases (such as autophagy in ALS and inflammatory bowel disease; insulin secretion in type 2 diabetes; or microglia activation in Alzheimer's);
- Create ***Tools to propel therapeutic development for high-priority targets*****,** to prioritize and validate mechanisms as therapeutic targets;
- Develop ***New methods for discovering biomarkers of disease*****,** to provide mechanistic insights and support clinical translation, including providing measures of target engagement and therapeutic response in clinical trials; and
- Accelerate ***Genomically-informed clinical trials*****,** to increase the success of trials while decreasing the number of patients needed, the number of adverse events, and the time to enrollment, by making greater use of genomic information.

Activities in these areas should focus on high-priority diseases, with a special focus on the Flagship Disease Projects (Recommendation 1).

We also discuss actions needed to support the clinical use of genomic information:

- Establish best practices for the ***Responsible use of polygenic scores in medicine***, including the creation, assessment, certification and clinical use of PRS's.

In addition to the recommendations above, ICDA should also explore how best to ensure that the development and implementation of genotype-stratified clinical care will benefit patients.

Recommendations and Implementation. We make five recommendations (Recommendations 10, 11, 12, 13, and 14). ICDA's activities related to these recommendations should be coordinated by the **ICDA Medicine Working Group** and the **ICDA Pharma Council**.

Recommendation 10: Develop functional assays for systematic dissection of disease-related processes

While general approaches (such as those described in the recommendations above) will provide an essential foundation for all diseases, the biological understanding and therapeutic development for specific diseases will also require:

1. rapid and efficient development of traditional and high-content high-throughput assays to help implicate genes and variants in cellular processes relevant to particular diseases (such as autophagy in ALS and inflammatory bowel disease; insulin secretion in type 2 diabetes; or microglia activation in Alzheimer's);
2. development of better cellular, organoid, tissue, and animal models in which to assay disease-relevant processes and environmental exposures; and
3. development of effective analytical methods to integrate information from diverse assays.

The scientific community should work together to develop and share such high-throughput assays for disease-relevant cellular processes, with a particular emphasis on the diseases in Flagship Disease Projects.

To advance this goal,

1. **ICDA should work with key stakeholders, including pharmaceutical companies, academic scientists, and funders, to:**
   a. **encourage systematic development of high-throughput functional assays for diseases of high interest for therapeutics;**
   b. **encourage systematic development of relevant cellular, tissue, and animal models for diseases of high interest for therapeutics;**
   c. **promote sharing of protocols, reagents, and results from these efforts.**
2. **Pharmaceutical companies and other funders should support efforts directed toward these goals.**

Particular priority should be given to diseases in the Flagship Disease Projects.

Recommendation 11: Accelerate therapeutic development of high-priority targets

Once a specific target has been identified (based on genome-scale data and disease-specific assays) as likely to be favorable for therapeutic development, there remains a further challenge: the function of the target is often poorly understood.

Characterizing the mechanisms through which a target gene or protein acts and validating its value as a therapeutic target requires having a range of tools, including gene-specific functional assays, antibodies, small molecule inhibitors and other tool compounds, structures, interaction maps, and biomarkers of target engagement. Creating these tools remains slow and laborious.

Efforts, such as the Structural Genomics Consortium, have been working to develop methods to substantially accelerate some of these steps and to apply these improved methods to high-priority targets.

Genomic approaches offer similar opportunities to substantially accelerate the ability to create the necessary tools to advance the targets into drug development.

To advance this goal, ICDA should partner with the Structural Genomics Consortium and other efforts to:

1. **develop genomic methods to improve the creation of functional information and reagents to advance targets into drug development, with a particular emphasis on the diseases in Flagship Disease Projects;**
2. **drive the application of these methods to high-priority targets emerging from genetic studies;**
3. **encourage the systematic development of efficient methods and capabilities to characterize the function of the "dark proteome" — that is, the majority of proteins about whose functions relatively little is known.**

Recommendation 12: Catalyze new methods for discovering biomarkers of disease

Biomarkers are increasingly essential for clinical development. Useful biomarkers can take a variety of forms, including proteins, metabolites and other analytes in plasma or other relevant biofluids, circulating DNA from liquid biopsies, gene expression in specific tissues or cell types, and various kinds of imaging. Such biomarkers, combined with genome-wide genetics, have the potential to both provide mechanistic insights and to support clinical translation—including acting as measures of target engagement and therapeutic response in clinical trials.

Despite their value, discovering useful biomarkers remains difficult and slow.

The ability to collect ever-increasing quantities and types of high-dimensional data (including multi-omic measurements, imaging data, rich medical records, and genetic associations with potential biomarkers, within and across populations) combined with advances in machine learning now offer the prospect of improving the systematic discovery of biomarkers. However, the scope of currently available datasets is modest. An important research goal will be to extend proteomic and metabolomic analysis (in plasma and/or serum initially at least) to both large population-based samples (such as national biobanks), and to other clinical investigation protocols including randomized controlled trials. There is a particular need to build datasets that track a range of omic phenotypes over time, across diverse ancestral backgrounds, and in relation to health, disease, and therapeutic response.

To advance this goal:

1. **Funders in academia and industry should encourage and support the application of existing and emerging approaches for the large-scale collection of biomarker data (in biobanks and clinical studies) across a range of modalities including those related to -omic measures (including proteomic and metabolomics), imaging and clinical monitoring (such as through apps and wearables) to define relationships between genetic variation, biomarker metrics, and disease state.**
2. **Funders in both academia and industry should encourage and support the development of new methods for biomarker discovery (such as those based on applying machine-learning methods) to complex multidimensional clinical and experimental data (including those derived from genomics, proteomics, imaging, clinical monitoring, and medical records).**
3. **ICDA should engage the scientific community to develop novel methods to discover and validate biomarkers, and to ensure global applicability of biomarkers.**

Identification of biomarkers will be critically dependent on many of the earlier recommendations, especially the collection and storage of biosamples gathered in large, diverse biobanks.

Efforts to use the explosion of high-dimensional data should engage the growing numbers of communities interested in the fusion of machine learning and biology—for example, the Bio-Turing Initiative at The Alan Turing Institute in the UK.

Recommendation 13: Advance Genomically-Informed Clinical Trials

There may be important opportunities to improve clinical trials, which are the most expensive element of drug development, by making greater use of genomic information. While genomic information is routinely used in testing cancer drugs, it is more rarely used in other areas of medicine.

Genomic information may potentially be useful in:

- decreasing the number of patients, by using genetic risk information (both monogenic mutations and polygenic scores) to select patients in which the background event rate is higher;
- increasing the success of clinical trials, by identifying and focusing on subsets of patients in which specific cellular and physiological mechanisms are affected (for example, distinct subgroups of disease identified based on gene-expression patterns in relevant cell types);
- decreasing adverse events, by using genomic information (such as through PheWAS and exploratory human trials) to anticipate and understand likely side effects and identify patients at greatest risk; and
- decreasing time to enrollment, by identifying potential patients through large biobanks with clinical and genomic information.

Considerable work, however, will be needed to validate whether and where genomic information might achieve these goals. Where genomic information is useful, it will also be important to determine how best to implement and support the use of such approaches in clinical trials—including working with regulatory agencies and ethics review boards.

Genomically-informed clinical trials would be most efficient if connected with biobanks assembled in the context of medical systems, where participants have consented to recall-by-genotype.

To advance this goal, ICDA should work with scientists in industry, academia, funders, and regulatory agencies to:

1. **explore how genomic approaches could best improve clinical trials; and**
2. **help develop and validate methods and best practices.**

In particular, ICDA should coordinate closely with the Joint Initiative on Good Practice in Clinical Research, established in 2019 by Wellcome Trust, the Gates Foundation, and the African Academy of Sciences, which is working to develop new clinical research guidelines to increase the efficiency of clinical trials.

Recommendation 14: Ensure Responsible Use of Polygenic Scores in Medicine

Polygenic risk scores (PRS), which aggregate genome-wide information to stratify individual risk of diseases in populations, have the potential to serve as biomarkers across a wide range of diseases and applications in research, clinical care, clinical trial design, and public health.

PRS will have the greatest positive impact in these applications when they are accurate, well-calibrated, and properly interpreted for the populations in which they are applied.

However, important challenges must be addressed regarding their generalizability across ancestries, demographic categories, and environmental exposures (which can affect heritability and their implementation in clinical practice (e.g., developing screening guidelines that combine PRS with traditional clinical risk factors).

To advance this goal, ICDA should establish a PRS Task Force that, in collaboration with other stakeholders, would:

1. **assess major gaps and risks;**
2. **establish best practices for the creation, assessment, certification, and clinical use of PRS;**
3. **address issues related to the applicability of PRS across and within ancestry, demographic groups (e.g., sex, age, and socioeconomic status), environmental exposures, and clinical risk factors; and**
4. **recommend actions to ensure that PRS are deployed responsibly and equitably, including that any PRS used in clinical practice is transparently disclosed.**

The PRS Task Force should include geneticists, clinicians, public health specialists, statisticians, ethicists, legal scholars, social scientists, and other relevant stakeholders to ensure that both scientific and societal issues are considered.

Direction 5: Data

It is essential that the scientific community be able to access and analyze the staggering amounts of data that already exist and will be generated in the years ahead. The scale is challenging. Whole-genome sequences have been generated for roughly one million individuals to date, with the total data storage corresponding to roughly 50 Petabytes; for comparison, the size of the Netflix movie corpus is only 3.14 Petabytes. Experience suggests that derived data types from various analyses may expand this number by a factor of 5-10x. Moreover, the volume of genomic data has been doubling nearly every 8 months, and this does not even account for the phenotypic data and other metadata.

As a result, our paradigms for genomic data sharing must be rethought. It is no longer feasible to download data to local servers for analysis. Instead of bringing data to the analysis, we need to bring analysis to the data, using modern cloud technology to enable streamlined data flow and harmonization in a federated environment. We also need to create facile user interfaces that allow non-computational biologists to utilize genomic data.

To accelerate progress in the M2M2M Challenge, we propose a focus on three important areas addressing data storage, access, harmonization, and use:

- Create a **Data Platform** to meet the data storage, access, harmonization, and analysis needs of the international community, while complying with international standards and policies;
- Develop **Genotype x Phenotype portals** to readily access information about the relationship between genes and phenotypes, including genetic data, functional assay data, environmental and exposure data, and clinical information; and
- Curate **Gold-standard validation datasets** for the robust and unbiased testing, benchmarking, and validating of novel computational methods to infer causality at all steps in the process from maps to mechanisms to medicine**.**

Early use-cases related to these activities should work across disease areas and the different components of the *M2M2M Challenge* to prevent the development of silos.

Recommendations and Implementation. We make three recommendations (Recommendations 15, 16, and 17). ICDA's activities related to these recommendations should be coordinated by an **ICDA Data Working Group**.

Recommendation 15: Create a Shared Data Platform

The scientific community, including data generators, researchers, and methods developers, needs a common infrastructural substrate to store, access, and use the massive datasets from large-scale studies of Maps and Mechanisms that already exist and the even larger datasets that will be generated in the years ahead.

A shared **Data Platform** should be developed to bring together these diverse user communities by offering a set of common features and functionalities for use across these groups. (A full list of important features and functionalities of such a Platform is outlined below.)

A shared Data Platform does not mean a single common database (in which all data reside in a single physical location), but a *federated computational ecosystem*, comprised of distinct data repositories designed to easily interact — more precisely, each being 'instances' built from two distinct components:

1. a repository environment for data storage and access, and
2. an analysis environment for computing across datasets to answer scientific questions.

The availability of a common Data Platform, based on open-source code and able to run on many clouds, would allow entities to create their own *distinct* secure and scalable Data Repositories for different purposes (from a large public repository freely available to users anywhere in the world, to a controlled-use data repository within a country, to a fully private data repository at a pharmaceutical company), while letting scientists (i) analyze their own data with shared, best-practice analytical tools and pipelines (as well as adding their own tools) and (ii) perform federated analyses *across* data repositories (to the extent permitted by applicable national regulations, IRB approvals, and data-use permissions).

The open-source nature of the software would also allow users to customize an individual instance to meet their unique needs, without losing the common software functionalities of the federated environment.

To advance these goals,

1. **ICDA should:**
   a. **define, in collaboration with funders, stakeholders and GA4GH, how a Data Platform could best serve the needs of the community—including researchers working on biobanks, human genetics, functional genomics, disease biology, clinical research, and therapeutic development, as well as ethicists and lawyers;**
   b. **encourage the development of a common Data Platform meeting the criteria above; and**
   c. **encourage the creation and operation of Data Repositories using the Data Platform.**
2. **Funders in both academia and industry should support:**
   a. **the development of a common Data Platform meeting the criteria above; and**
   b. **the creation and operation of Data Repositories using the Data Platform.**

A Data Platform should meet several criteria to support human genetics research:

- easily handle large-scale human genetics and functional genomics datasets, (including multiple different layers of raw, processed, and normalized data; metadata about protocols used for experimental data collection; and standards and formats for sharing key types of functional data and biological results);
- enable cloud-based storage and access to massive genomic datasets, with the ability to run on multiple cloud platforms, including the ability to easily use shared controls in association analyses;
- provide data processing and analytical software (such as imputation for genotype array data, variant calling for sequence data, and quality control pipelines) that can run seamlessly on the data in the platform;
- support integration, analysis, and visualization of multiple data types (e.g. genomics, transcriptomics, proteomics, metabolomics, social factors, exposure and environmental factors (including physical, chemical, biological, and built environment), diet and lifestyle, imaging, clinical phenotypes, and more);
- support interoperable usage, including storage, analysis, and visualization, across individual and summarized data from multiple data sources (e.g. repositories built in different countries and focusing on different populations);
- enable flexible access to data, including easy creation of data portals, to give users easy views of results and insights to meet a wide range of needs from clinical interpretation to biological explorations;
- enable users to use their own pipelines and software;
- enable collaborative and reproducible research, so researchers in different organizations can build on each others' efforts;
- be compatible with ongoing efforts to extract harmonized phenotypes from medical records;
- have the ability to manage user approval and authentication, to ensure that only authorized researchers may gain access;
- incorporate an electronic consent management system that allows datasets to specify consent categories (that is, well-structured secondary data use restrictions that address privacy and confidentiality, data sharing, return of individual research results, protecting the interests of communities and populations, and commercialization of research, etc.) and automatically grants access to research projects that are consistent with these consent categories;
- facilitate genomics research while protecting the rights and welfare of human subjects;
- ensure security and comply with regulatory requirements, including data locality requirements;
- be compliant with GA4GH standards;
- conform to the Data Biosphere principles -- modular, community-driven, open, and standards-based; and
- align with FAIR principles for scientific data management and stewardship.

Recommendation 16: Create Genotype x Phenotype Portals

As the types and amounts of information grow, it is increasingly important to have coherent ways to view *all* information about the relationships between genes and phenotypes derived from human genetics, functional genomics and more.

The scientific community will need portals that make it easy for anyone to fetch and integrate many different kinds of information from data repositories, including:

- Allele frequency references (e.g., gnomAD or BRAVO);
- PheWAS browsers (e.g., pheweb) to display all associations for a given variant or gene;
- Polygenic Risk Score repositories (with pre-calculated weights);
- Dose-response server with curated allelic series for gene dosages;
- Human knock-outs linked with phenotypes to assess the impact of complete gene loss;
- Coding variant portals;
- Molecular assay portals (e.g., GTEx portal); and
- Disease-specific information integrating all phenotypes for an outcome including susceptibility, severity, and age of onset.

Moreover, different portals will need to serve different purposes — such as discovering disease genes, inferring the cell types and pathways in which a gene acts, clinical interpretation of variants, and evaluating a potential therapeutic target.

To advance this goal,

1. **ICDA should work with key stakeholders and funders to:**
   a. **scope the scientific needs for such portals; and**
   b. **develop a plan to ensure the creation of such portals.**
2. **Funders should support the creation of such portals.**

Recommendation 17: Create Gold-Standard Validation Datasets

Solving the *M2M2M Challenge* will benefit from shared, well-curated validation datasets.

1. To advance the development of computational tools for data processing, we need deeply characterized, readily shareable genomic benchmark datasets — for example, to demonstrate consistency or functional equivalence of processing pipelines, evaluate base error rates and phasing quality, and compare the performance of algorithms.
2. To develop and validate reliable and robust computational methods to identify key steps along the causal chain (from associated variants to causal variants, genes, transcripts, proteins, biomarkers, endophenotypes, physiology, and pathophysiology), we need shared reference sets of “gold standard” results (with both “true positives” and “true negatives”) that are as unbiased and as comprehensive as possible, and that include the nature and degree of supporting evidence. Examples include: confirmed disease alleles, credible sets at a GWAS locus, validated connections

between enhancers and their target genes (based on CRISPR-based gene disruption), and effects of coding variants on cellular programs (based on isogenic editing).

These gold standard reference sets should be curated based on community-driven definitions.

To advance this goal,

1. **ICDA should**
    a. **work with experts to define suitable criteria for genomic benchmark datasets for validating computational pipelines;**
    b. **work with experts to define suitable criteria for evidence of 'causality' (e.g., of variants, genes, transcript, etc.);**
    c. **develop a plan to ensure the creation, curation, and maintenance of a suitable repository for gold standard gene sets and the necessary meta-data; and**
    d. **champion and encourage submission, curation, and deposition of suitable gold-standard gene sets.**
2. **Funders should support well-conceived and well-planned efforts to develop and maintain such gold-standard reference sets.**

ICDA should engage with the growing numbers of stakeholders interested in the establishment of these benchmark criteria and accompanying computational pipelines, such as Open Targets in the UK.

Direction 6: Policy

Data sharing has played an essential role in driving progress to study the genetic basis of diseases, from our knowledge of the human genome to clinical practice. It is therefore essential to have carefully crafted policies that promote data sharing, protect participants, comply with national regulations, and respect cultural perspectives.

While there has been progress in developing appropriate policies, recent experience has revealed important gaps that urgently need to be addressed—by bringing together stakeholders to clarify or improve current policies and to develop new policies where needed.

These stakeholders should include a wide range of expertise and perspectives.

Recommendations and Implementation. We make one recommendation (Recommendation 18). ICDA's activities related to these recommendations should be coordinated by an **ICDA Data Compliance Task Force** within the ICDA Data Working Group. Activities should be closely coordinated with **GA4GH**.

Recommendation 18: Develop and Inform Policies for Genetic Studies, Data Sharing, and Federated Analysis

Scientific and medical progress would be accelerated by improved policies around consent, access, and sharing of genomic and clinical data. Some examples include:

- Policies that would allow automatic determination of which investigators may use which datasets, based on (i) standardized "consent categories" that classify datasets based on permitted uses (e.g., general research use, health related outcomes, etc.) and restrictions on secondary use (e.g., restrictions on commercial use or limitation to disease-specific uses); (ii) standardized "user categories," mirroring the consent categories, that classify research projects according to their approved uses (e.g., a project on genetic discovery in schizophrenia would be granted access to any dataset that is classified under general research use, health-related outcomes, or schizophrenia-specific research).
- Policies to ensure the deposit of GWAS summary statistics and associated metadata (for example, as a requirement for publication and grant renewal);
- Policies (and supporting technologies) to facilitate federated analysis of individual-level data held in different jurisdictions, respecting national regulations concerning data sharing;
- Policies that make clear when aggregated data may be treated as open-access vs. controlled access;
- Policies concerning the use of genomic data without phenotype information to strengthen genetic resources (e.g., for imputation, allele frequency estimation, or use as population controls); and
- Guidelines for effective regional and local community engagement, to ensure inclusive policies and culturally acceptable research practices, especially for how samples are used (e.g., for creating cell lines or *in vivo* models).

To advance this goal, ICDA, in collaboration with global stakeholders, should:

1. **establish a Data Compliance Task Force to review the policy needs and develop concrete proposals; and**
2. **work with national authorities and funders to refine and implement these proposals.**

The Data Compliance Task Force should have diverse expertise and representation, to ensure that proposed policies reflect global needs and perspectives. It should coordinate closely with GA4GH.

Direction 7: Global Equity

Achieving the ICDA's goals will require global engagement of *both* research participants and scientific researchers. Indeed, the two are inextricably linked: the lack of diversity among the former is due in large measure to the lack of full, fair, and equitable inclusion among the latter.

There are many serious barriers to full inclusion among scientific researchers related to training, travel, facilities, computational infrastructure, pricing of equipment and reagents, and research funding.

ICDA should work to lower these barriers, by working with regional participants to understand the barriers, develop effective solutions, and implement them.

Recommendations and Implementation. We make one recommendation (Recommendation 19). ICDA's activities related to this recommendation should be coordinated by the **ICDA Organizing Committee**.

Recommendation 19: Promote Global Equity

To realize its goals and to serve the world, the mission of understanding and treatment of common diseases must be a global activity — involving not only research participants across the globe, as noted above, but also scientific researchers across the globe. Indeed, the two are inextricably linked.

Yet, there are many barriers to conducting world-class projects in many regions — including access to training, technology and funding, and to the benefits of the research.

It is critical to take active steps to decrease these barriers.

To advance this goal,

1. **Funders should support well-conceived training programs aimed at promoting global equity in the field — potentially ranging from undergraduate internships in world-class laboratories to postdoctoral fellowships to conduct long-term projects and then implement new research lines in their home countries, as well as hands-on training workshops and 'hackathons'; and**
2. **ICDA should work to secure equitable access to advanced technologies needed to tackle the M2M2M Challenge (such as genome sequencing, single-cell characterization, and computational analyses) — including by:**
   a. **negotiating favorable pricing for equipment and reagents to facilitate global participation;**
   b. **promoting the development of online tutorials and courses, to ensure access to knowledge for scientists and trainees with limited ability to travel; and**
   c. **ensuring that data resources are designed to accommodate settings with a range of communications bandwidths.**

Direction 8: Innovation and Training

While the recommendations above relate to larger efforts that need to be undertaken, we strongly believe that individual investigator-initiated projects to pursue new analysis methods, new technologies, and new biological directions will also be critical for tackling the M2M2M Challenge. Additionally, training programs to help develop new investigators will be crucial.

Recommendations and Implementation. We make four recommendations (Recommendations 20, 21, 22, and 23), which are primarily directed to funders.

Recommendation 20: Support Development of New Analysis Methods

There will be a tremendous need for new analysis methods, to match the explosion of new data generation. It is critical to support creative efforts to develop new analytical methods. Examples of methods that are needed by the community include:

- Methods for comprehensive phenotype definitions from longitudinal and comorbidity data;
- Methods to improve genetic mapping by, for example, improved imputation, improved correction for confounding, and improved ways to leverage multiple related phenotypes and pedigree information to gain power;
- Methods to infer and predict causal roles for specific variants, genes, cell types, and pathways, based on genetic and functional data, and to partition these roles into distinct mechanisms;
- Methods to link generic cellular phenotypes (e.g., scRNA-seq) to physiologically relevant cellular functions (e.g., autophagy);
- Methods to incorporate environmental exposures (both genotoxic and non-genotoxic) into functional and mechanistic studies, including environmental perturbation studies;
- Methods to accelerate *in silico* screening of novel therapeutic and drug targets by utilizing large-scale human genotype, epigenome, and phenotype association resources;
- Methods to better translate clinical information across populations, including admixed populations—including methods to characterize shared and non-shared genetic components of disease risk across populations and methods to predict disease risk that translate effectively across populations.

To advance this goal,

1. **Funders should support a robust portfolio of investigator-initiated grants (for individuals and collaborative teams) to develop new methods relevant to the M2M2M Challenge.**
2. **ICDA should highlight such work at its meetings.**

Recommendation 21: Support Development of New Technology

There will be a tremendous need for creative new experimental technologies, to connect maps to mechanisms to medicine.

Among many possible examples are:

- High-throughput methods to introduce variants into the genome at high accuracy and characterize their effects on chromatin state, gene expression, and cellular functions;

- High-throughput methods to perturb genes and measure their downstream effects on cellular programs;
- High-throughput molecular and cellular phenotyping assays, in bulk and in single cells, based on sequencing assays, imaging, and proteomics;
- Informative and scalable cellular, organoid, and animal model systems for different diseases; and
- Efficient and high-throughput methods for measuring environmental exposures and for integrating the environment and environmental exposure into genomic studies.

To advance this goal,

1. **Funders should support a robust portfolio of investigator-initiated grants (for individuals as well as collaborative teams) to develop new technologies relevant to the M2M2M Challenge.**
2. **ICDA should highlight such work at its meetings.**

Recommendation 22: Support Research in New Biological Directions

While there has been much progress concerning the principles underlying the genetic basis of common disease, there are likely to be many fundamental surprises ahead.

Some interesting possible examples include a greater role for somatic mutations in a range of common diseases and interactions between genetic and environmental risk factors.

It is critical to support creative efforts to explore new biological directions.

To advance this goal,

1. **Funders should support a robust portfolio of investigator-initiated grants (for individuals and collaborative teams) to explore new biological directions relevant to the M2M2M Challenge.**
2. **ICDA should highlight such work at its meetings.**

Recommendation 23: Support Establishment of New Training Programs

A generation of new investigators with scientific perspectives and skills spanning from maps to mechanisms to medicine will be a critical foundation for achieving the long term goals.

It is therefore critical to have robust training programs for early-career scientists.

To advance this goal,

1. **Funders should support a robust portfolio of relevant training programs — including interdisciplinary, cross-disciplinary, and skill-specific training; and**
2. **ICDA should help promote training opportunities, including workshops on relevant topics.**

IV. Conclusion

The ICDA Recommendations (v1.0) reflect the input of a diverse and engaged community of researchers, clinicians, funders, and policy experts. We expect that the Recommendations will continue to evolve and mature with progress in the field, based on input from the scientific community. We welcome your input on how we might continue to refine and improve these recommendations at https://www.icda.bio/.

The next step for ICDA will be to work with the scientific community to help implement these recommendations.

V. Contributors

ICDA Organizing Committee

Co-Chairs

Eric Lander	Broad Institute of MIT and Harvard
Cecilia Lindgren	University of Oxford

Executive Director

Rachel Liao	Broad Institute of MIT and Harvard

Members

Søren Brunak	University of Copenhagen
Judy Cho	Icahn School of Medicine at Mount Sinai
Rory Collins	University of Oxford
Nancy Cox	Vanderbilt University
Mark Daly	Institute of Molecular Medicine Finland (FIMM), University of Helsinki
George Davey Smith	University of Bristol
Emmanouil Dermitzakis	University of Geneva
Michael Dunn	Wellcome Trust
Lude Franke	University Medical Centre Groningen
David Glazer	Verily Life Sciences
Matthew Hurles	Wellcome Sanger Institute
Carolyn Hutter	National Human Genome Research Institute
Nancy Ip	Hong Kong University of Science and Technology
Sally John	Biogen
Tuuli Lappalainen	New York Genome Center, Columbia University
Partha Majumder	National Institute of Biomedical Genomics India
Mark McCarthy	Genentech
Andrés Moreno Estrada	National Laboratory of Genomics for Biodiversity Mexico
Benjamin Neale	Broad Institute of MIT and Harvard, Massachusetts General Hospital
Yukinori Okada	Osaka University Graduate School of Medicine
Helen Parkinson	EMBL-European Bioinformatics Institute
Charles Rotimi	National Human Genome Research Institute
Jay Shendure	University of Washington / Howard Hughes Medical Institute
Nicole Soranzo	Wellcome Sanger Institute

Kári Stefánsson	deCODE genetics
Patrick Sullivan	University of Northern Carolina, Karolinska Institutet
E Shyong Tai	National University of Singapore
Nicki Tiffin	University of Cape Town
Ricardo Verdugo	University of Chile
Cristen Willer	University of Michigan
Ambroise Wonkam	University of Cape Town
Unnur Þorsteinsdóttir	deCODE genetics

Maps Working Group

Co-Chairs

Benjamin Neale	Broad Institute of MIT and Harvard, Massachusetts General Hospital
Cristen Willer	University of Michigan

Members

Mark Daly	Institute of Molecular Medicine Finland (FIMM), University of Helsinki
Meg Ehm	GlaxoSmithKline
Ira Hall	Washington University in St. Louis
Eimear Kenny	Icahn School of Medicine at Mount Sinai
Iscia Lopes-Cendes	Brazilian Initiative on Precision Medicine - BIPMed
Partha Majumder	National Institute of Biomedical Genomics India
Sarah Medland	QIMR Berghofer Medical Research Institute
Andrés Moreno Estrada	National Laboratory of Genomics for Biodiversity Mexico
Yukinori Okada	Osaka University Graduate School of Medicine
Manny Rivas	Stanford University
Charles Rotimi	National Human Genome Research Institute
Danish Saleheen	Columbia University
Pak Sham	Centre for Genomic Sciences, The University of Hong Kong
Xueling Sim	National University of Singapore
Nicki Tiffin	University of Cape Town
Ricardo Verdugo	University of Chile
Helen Warren	Queen Mary University of London

Mechanisms Working Group

Co-Chairs

Tuuli Lappalainen	New York Genome Center, Columbia University
Jay Shendure	University of Washington, Howard Hughes Medical Institute

Members

Nadav Ahituv	University of California San Francisco
Melina Claussnitzer	Broad Institute of MIT and Harvard, Beth Israel Deaconess Medical Center
Jesse Engreitz	Broad Institute of MIT and Harvard
Eric Fauman	Pfizer
Lude Franke	University Medical Centre Groningen
Anshul Kundaje	Stanford University
Kasper Lage	Massachusetts General Hospital
Bogdan Pasaniuc	University of California Los Angeles
Tune Pers	Boston Children's Hospital
Nicole Soranzo	Wellcome Sanger Institute
Fabian Theis	Helmholtz Zentrum München
Gosia Trynka	Wellcome Sanger Institute

Medicine Working Group

Co-Chairs

Judy Cho	Icahn School of Medicine at Mount Sinai
Mark McCarthy	Genentech
Unnur Þorsteinsdóttir	deCODE genetics

Members

Melina Claussnitzer	Broad Institute of MIT and Harvard, Beth Israel Deaconess Medical Center
Anna Gloyn	Stanford University
Nancy Ip	Hong Kong University of Science and Technology
Yukinori Okada	Osaka University Graduate School of Medicine
Danielle Posthuma	Vrije Universiteit (VU) Amsterdam
Nadeem Sarwar	Eisai
E Shyong Tai	National University of Singapore
Olga Troyanskaya	Princeton University, Flatiron Institute of the Simons Foundation
Ambroise Wonkam	University of Cape Town

From Maps to Mechanisms to Medicine:

Using human genetics to propel the understanding and treatment of common diseases

International Common Disease Alliance
White Paper v1.0

Developed by the ICDA Organizing Committee
(See contributor list at end)

INTERNATIONAL COMMON DISEASE ALLIANCE | Maps to Mechanisms to Medicine

BLANK PAGE

Preamble

Human genetics stands at a pivotal moment. The past several decades have seen enormous progress. Various efforts—the Human Genome Project, deep catalogs of genetic variation, powerful study designs and new analysis methods—have resulted in a wealth of knowledge about the genetics of common disease. This includes the discovery of more than 70,000 robust genetic associations to common diseases and traits, and important insights into the underlying basis of some diseases. Yet, there is much more to be done.

It is time to envision the next phase of human genetics—to accelerate progress in moving from Maps to Mechanisms to Medicine (we have termed this the *M2M2M Challenge*).

Achieving the promise will require deep collaboration among many communities and stakeholders from across academia, medicine, biopharma, tech companies, and funders. The International Common Disease Alliance (ICDA) was formed to serve as a scientific forum to bring together key stakeholders from the global community to (i) identify common barriers to progress across diseases and populations, (ii) develop and promote collaborative solutions to overcome these barriers, and (iii) bring together the scientific community on an ongoing basis to share results, assess progress, and update plans.

The purpose of this ICDA White Paper is to map out the many pieces—the key knowledge, comprehensive datasets, new technologies, new analytical methods, computational platforms, data sharing frameworks, mechanistic assays, and drug development approaches—that need to be filled in to propel the next phase of human genetics. The goal is not a monolithic project, but a vibrant ecosystem.

The ICDA White Paper is intended as a living document. An initial version (v0.1) was released in September 2019, to provide a framework for a conversation among the entire ICDA community.

Chapter 1 summarizes progress to date in the genetics of common diseases and defines the challenge ahead in moving rapidly from Maps to Mechanisms to Medicine. (We underscore that the text is only a summary: it does not provide a comprehensive review and does not currently contain citations to the literature.)

Chapter 2 looks to the next phase of human genetics, framing some overarching goals and describing examples of foundational resources that may be needed.

ICDA Recommendations. The purpose of this ICDA White Paper is to look back at the progress over the past two decades, identify the challenges ahead and share an initial vision for the next phase of human genetics.

ICDA's subsequent step was to develop a framework and specific recommendations to address the M2M2M Challenge, through key projects, platforms, resources and policies.

ICDA released Draft Recommendations (v0.9) in February 2020 and a first full version of Recommendations (v1.0) in July 2020.

Chapter I. Progress to Date and Challenge Ahead

Common diseases — those affecting more than one in a thousand individuals — account for the majority of human morbidity and mortality, and they represent major healthcare challenges around the world. Most of these diseases lack effective therapies that benefit all patients. Despite huge investments, drug development for common diseases is expensive, slow and prone to failure, often due to lack of efficacy. Rapid development of optimal therapies for common diseases is thus a primary challenge facing biomedical research in the 21st century.

For almost all common diseases, including non-communicable diseases and infectious diseases, genetics plays a substantial role in disease susceptibility, disease progression, and/or response to therapies. It is becoming increasingly clear that genetics offers a powerful approach to propel the understanding and treatment of common diseases.

1. Unique role of genetics in medicine

Genetics plays a unique role within medicine, because it provides a way to discover the causal biological mechanisms of any disease with no prior biological hypothesis about the cell types or processes involved. Every other biological observation associated with a disease might either be a cause or an effect. For genetic variation, the arrow of causality runs in only one direction: from genotype to phenotype.

Advances in genetics now make it possible to comprehensively screen genetic variation in patients with common diseases to enable:

- **systematic discovery and characterization of disease mechanisms**, which remain poorly understood for most common diseases and are crucial for biomedical progress.
- **identification of novel drug targets,** as well as prioritization of existing drugs and drug targets for accelerated development. Retrospective studies of drug development pipelines have shown that robust genetic evidence is predictive of success in clinical trials—prompting pharmaceutical companies to invest substantial resources in ensuring that their decision-making is grounded in genetics whenever possible.
- **preventative medicine** to allow individuals with a genetic predisposition to a particular common disease to be identified earlier in life, helping them to decrease their risk by means of life-style changes (e.g., altering diet, increasing exercise, losing weight, or stopping smoking), medication (e.g., taking statins to lower risk of cardiovascular disease), screening (e.g., earlier mammography or other cancer screening, cholesterol levels, imaging), or perhaps preventive surgery (e.g. repair of aortic aneurysm or heart valves). Because our genetic makeup is set at conception, genetics has a unique potential for predicting disease risk, avoiding disease, or treating it early.
- **precision medicine** (or stratified medicine) to target therapies that can be transformative in some patients, while ineffective in others (e.g. anti-TNF therapies in rheumatoid arthritis), by identifying which patients are likely to receive benefit and reducing harms from side-effects in patients unlikely to benefit.

2. The genetic basis of human diseases: What we have learned so far

2.1 Mapping rare monogenic diseases.

Geneticists have long known that many rare diseases (frequencies in the range 10^{-4}-10^{-6}) were transmitted in families in a manner consistent with a Mendelian, single-gene pattern of inheritance. However, there was no way to discover the genetic basis of the diseases without precise prior biological knowledge (e.g. that hemoglobin was disrupted in sickle cell anemia).

The idea that disease genes could be *systematically* discovered was first proposed for rare Mendelian diseases by Botstein and colleagues in the 1980s. They realized that the classical principles of genetic linkage used to map the location of mutations in experimental crosses in fruit flies and yeast could be applied to humans—provided one had a genetic map of common variants to trace the inheritance of chromosomal loci in families. Once linkage mapping revealed the disease locus, the gene could presumably be found by looking for rare mutations in the DNA sequence of nearby genes.

Making this vision feasible would require transforming our ability to analyze the human genome. It sparked the international Human Genome Project. The scientific community rallied around this ambitious intellectual project—developing ideas, technology and infrastructure, and making them freely available to all. Launched in 1990, it reached completion around 2003.

With the foundations in place, human geneticists have applied modern technologies for genotyping and sequencing to affected families and cases to identify the genes underlying more than 5,000 Mendelian disorders. The rare diseases that have been mapped are almost always caused by rare mutations of strong effect in the coding regions of individual genes. While individually rare, Mendelian diseases collectively affect roughly 1 in 100 live births. Knowledge of the genes has transformed genetic diagnostics and is propelling therapeutic efforts for rare diseases—driven by small-molecule drugs (such as cystic fibrosis), antisense oligonucleotides, gene therapy, and the prospect of genome editing.

***2.2 Common diseases have a different genetic architecture*.**

The vast majority of human morbidity and mortality, however, is due to *common* diseases—such as cardiovascular disease, diabetes, cancer, schizophrenia, Alzheimer's disease, autism, inflammatory bowel disease and many more conditions.

From the outset, human geneticists wondered how to bring genetic mapping to bear on common disease. Many clues hinted that the genetic architecture of most common diseases and traits was polygenic —that is, they were shaped by many variants affecting many genes. Quantitative trait locus mapping in plants and animals confirmed longstanding ideas about the polygenic nature of common traits. Genetic epidemiology found that common disease risk did not show the sharp fall-off with decreasing relatedness expected for a monogenic etiology. And, most importantly, traditional family-based linkage mapping studies of common diseases in the 1990s revealed only a few reproducible loci in a handful of common diseases and none in most—despite a much greater scale than had been applied to the parallel studies of rare disease. Furthermore, the distinctive history of human populations—a small initial size followed by rapid exponential expansion—suggested that the allelic frequency spectrum of common diseases should be different than for Mendelian diseases, with common variants playing a much larger role in common diseases. By the turn of the century, the challenge was clear: to develop a paradigm (analogous to linkage mapping for Mendelian diseases) to associate genetic variants with common disease.

2.3 Discovering common variants underlying common diseases: From families to populations.

The solution was to go beyond the classical principles of tracing genes in *families* by linkage mapping to create new principles to trace them across *populations* by genome-wide linkage-disequilibrium mapping — first in isolated populations such as Finland, then extended to any human population. Linkage-disequilibrium mapping exploited the fact (again due to the recency of human expansion and very low mutation rate of human DNA) that most common variants in a population reside on common ancestral segments of DNA that can be detected by using a very dense genetic map. By comparing the frequency of these short ancestral segments in cases and controls for a common disease or trait, one can infer if the ancestral segment carries an allele that alters disease risk. The approach works regardless of whether the disease is monogenic or polygenic.

The approach—that is, scanning the genome with a dense collection of common genetic variants to detect association between individual variants and a phenotype, and thereby regions of linkage disequilibrium containing causal genes—is typically referred to as a Genome-Wide Association Study (GWAS). For a given phenotype, GWAS assesses the degree to which each genetic variant across the genome is correlated with the phenotype and thereby identifies *disease-associated loci*. (Mathematically, GWAS yields a *genome-wide vector of effect sizes*, showing the marginal effect of each variant, and a p-value, reflecting the probability that the observed effect would occur by chance. Disease-associated loci are defined as regions where the p-values meet stringent criteria for genome-wide significance.)

GWAS focuses on *common* variants, defined in the operational sense that they occur frequently enough that study sizes typically provide reasonable statistical power to test association for each variant *individually*. For modern studies with tens of thousands of participants, variants can be regarded as common for the purpose of analysis if they have allele frequency of ~0.1% (and, in some cases, even lower).

GWAS are sometimes called Common Variant Association Studies (CVAS) to distinguish them from complementary methods called Rare-Variant Association Studies (RVAS) that are increasingly being applied to detect association with rare variants.

CVAS and RVAS differ in two ways. The first difference is methodological and more fundamental. Whereas common variants can be assessed individually, analyzing rare variants requires deciding which ones to aggregate together. As discussed below, this can be challenging. The second difference is technological and temporary. Whereas common variants can be catalogued in advance and assayed (directly, or indirectly by imputation) with inexpensive genotyping arrays, the set of rare variants in a sample must be detected by direct sequencing, which is more costly. For both reasons, RVAS has so far largely focused on rare variants in coding regions.

At a high level, though, the basic principle is the same: to look for genetic differences between people with disease and people without disease. As the cost of whole-genome sequencing declines, we will increasingly analyze both common and rare variants together to obtain a more complete picture of the genetic architecture.

2.4 Common Variant Association Studies.

A tremendous amount has been learned from GWAS with common variants since these studies began more than a decade ago.

Enabling GWAS: foundations. Just as for discovering genes that cause Mendelian diseases, a major scientific paradigm was needed to enable genetic studies of common diseases. Starting in the mid-1990s, the work included developing (i) an initial understanding of the genetic structure of the human population, including the extent of variation and linkage disequilibrium in various populations; (ii) a near-complete catalog of the tens of millions of common genetic variants in the human population and a high-resolution map of the haplotype structure of the human genome (through international collaborations such as the SNP Consortium, the Haplotype Map Consortium, 1000 Genomes Project and Haplotype Reference Consortium); (iii) affordable methods for assaying variants—using genotyping arrays able to assay up to one million variants, coupled with highly effective imputation methods to infer most of the rest; (iv) methods to detect and correct for population substructure (ancestry differences between cases and controls); (v) rigorous thresholds for statistical significance, to ensure reproducible results; and (vi) open-source analysis software and data-sharing to empower the research community.

Enabling GWAS: power of large sample collections. The final component was an understanding of the scale necessary for success. Early GWAS studies were regarded by many as disappointing, because they yielded only a handful of loci that together explained only a tiny fraction (often <1%) of disease heritability —raising concerns about 'the mystery of missing heritability'. In fact, the problem was that these early studies, which typically involved only about a thousand cases, were seriously underpowered. Whereas an initial GWAS of schizophrenia with ~3,000 cases and ~3,000 controls identified no statistically significant loci, the most recent study (PGC Schizophrenia Working Group), with ~67,000 cases and ~94,000 controls, reveals 245 loci. Table 1 illustrates current progress for a few diseases and traits—resulting from tireless work by large groups of scientists for over a decade—to bring together hundreds of thousands of study participants, assemble their phenotypes, genotype their DNA, and statistically analyze the resulting data.

Disease	**Loci Mapped**
Type 2 Diabetes	403
Inflammatory bowel disease	273
Coronary artery disease	166
Schizophrenia	245
Rheumatoid arthritis	101
Obesity (BMI)	941
Height	3290
Fat distribution	463

Table 1. Number of loci to date associated at genome-wide significance, for several well-studied common diseases and traits.

It is now clear that at least hundreds and likely thousands of common variants contribute to most complex traits. Nearly all have only modest effects on disease risk, but they shed light on the disease mechanism and can identify important drug targets. For example, a common non-coding variant affecting the gene encoding HMGCoA reductase affects LDL-cholesterol levels by < 4%, but drugs that target the enzyme (statins) dramatically reduce LDL levels (by ~50%) and heart-attack risk.

The majority of GWAS loci reside in non-coding regions, and they are highly enriched in regulatory regions, such as cell type-specific enhancers, and splice sites. The minority that reside in protein-coding regions have, on average, somewhat larger effect sizes. Based on statistical analyses, the additive effects of the variants captured in current studies likely account for more than half of the heritability, as estimated from epidemiology. The remainder of the estimated heritability may be due to non-additive interactions among these loci (although there is currently no power to detect them) and rare variants. Additionally, epidemiological methods may overestimate the true heritability for various reasons, including shared family environment and genetic heterogeneity (where two or more distinct disease entities are combined).

The necessity of large sample sizes has sparked the formation of disease-focused consortia, pooling data from patients and controls both within and across countries for psychiatric diseases, cardiovascular diseases, metabolic diseases, autoimmune diseases, various cancers and many more conditions. Across all diseases, many millions of individuals have been analysed by genome-wide genotyping. Each disease study can ideally contribute information to the others, including serving as population controls.

In addition to creating disease-focused consortia, the human genetics community has also launched large biobanks that assemble a cross-section of the population with broad phenotypic characterization, typically including medical records and participant responses to a vast range of survey questions. Pioneering work at deCODE genetics in Iceland was an early exemplar. Subsequently, large biobanks have been or are being created by various other nations — efforts such as the UK Biobank (~500,000 participants), the FinnGen Project (currently ~180,000, with plans to reach 500,000), the Estonian Biobank (~150,000), BioBank Japan Project (~180,000), and the US All of Us Project (getting underway, with a target of one million) — and by some major medical centers (with early leaders such as Vanderbilt's BioVU, Geisinger's MyCode and Mt Sinai's BioMe).

Because the biobanks are not focused on specific diseases, they are currently underpowered for GWAS analysis of most diseases. A random sample of the population would be expected to contain only a limited number of cases of any particular disease, and current biobanks tend to be somewhat skewed toward healthier individuals. For example, a disease with 1% prevalence would have only 5000 cases in the UK Biobank. Some diseases, though, are greatly underrepresented — for example, schizophrenia has ~1% prevalence but fewer than 400 cases in the UK Biobank. Neurodegenerative and ophthalmologic diseases of aging are similarly underrepresented. Biobanks, however, are ideal for studying quantitative traits, very common phenotypes, and highly prevalent diseases (such as diabetes, present in ~6-10% of adults in Western populations).

As the aggregate sample sizes of biobanks climb into the many tens of millions, it should eventually be possible to study *any* common disease or trait — provided appropriate ways are developed to allow joint analyses across disparate biobanks, including comparable and robust standards for phenotyping. With access to detailed, longitudinal, harmonized medical records, it should also be possible to study not only disease incidence, but also disease progression, and response to treatments.

Enabling G x E: Genetic variants can interact with environmental exposures to affect biological pathways implicated in many common complex disease outcomes. Gene-environment (G x E) interaction studies play an important role in identifying environmental risk factors that may alter disease onset, progression, or severity. (Environment refers to a wide variety of exposures, including dietary, pharmacologic, socioeconomic, behavioral, occupational, infectious, chemical, and physical exposures.) Environmental factors may be modifiable and thus provide powerful strategies for disease prevention.

2.5 Rare Variant Association Studies.

Rare variant association studies could not begin in earnest until sequencing costs dropped sufficiently, but they have begun to contribute significantly to our understanding of common diseases. By RVAS, we refer particularly to studies of variants whose frequency is so low that they cannot be tested for association individually. Instead, the rare variants must be aggregated to test for differing frequencies in cases vs. controls.

Searching for disease association is more challenging and more expensive for rare variants than for common variants. First, the rare variants cannot be cataloged in advance but must be identified by sequencing each patient. Second, we must decide how to aggregate them. While straightforward for coding regions, this question is much trickier for the rest of the genome.

Early ‘rare variant studies’ involved some variants that, while they then required sequencing to identify, would today be called common. An important example was a stop codon in *PCSK9* (at 2% frequency in African Americans) that substantially lowers LDL cholesterol and was later shown to reduce risk of heart disease. Gene-based sequencing of several lipid-related genes (such as *ANGPTL3*, *APOB*, and *LDLR*) also found an excess of rare and low frequency variants in patients with heart disease. These studies encouraged efforts to undertake systematic RVAS in coding regions.

RVAS in coding regions. While common variants with modest effects likely explain more than half of the estimated heritability of common diseases, rare variants also play an important role. Rare variants are enriched for alleles that have larger effects on disease risk (because such alleles are typically prevented by natural selection from reaching higher frequency). Moreover, rare variants of large effect appear to be enriched in coding regions.

Rare coding variants are particularly valuable because (i) they provide crucial information about the direction of effect (e.g., whether loss-of-function (LoF) increases or decreases disease risk), (ii) they facilitate testing of protein function in model systems, (iii) they are well suited for direct physiological studies in human patients; and (iv) they may point to genes and drug targets that are less readily found by GWAS. (Conversely, some genes that can be identified by GWAS may not contain any associated rare variants that can be discovered by RVAS—because coding alterations affect the protein in all cells in which a transcript is expressed and may seriously disrupt protein function, whereas regulatory changes are often tissue specific and milder.)

RVAS in coding regions: power. To date, RVAS has largely focused on coding regions, where the effects of genetic variants on a transcript can be sensibly categorized to allow variant aggregation. In particular, it is possible to recognize many alleles that are predicted to cause truncation and thereby loss of a gene copy by virtue of altering a stop codon, creating frameshift or affecting a splice site—collectively termed LoF below. (These variants are also called protein-truncating variants (PTVs) or likely gene-disrupting variants.)

The sample sizes required for RVAS in coding regions are now well understood, based on (i) power calculations, conditional on the frequency and effect size for LoF alleles in a gene, and (ii) large databases (e.g., ExAC and gnomAD) that provide empirical frequencies for LoF frequencies for each human gene. The frequency of LoF alleles varies by two orders of magnitude across genes, with the 25th, 50th and 75th percentiles corresponding to 10, 1.5, and 0.6 LoFs per 10,000 chromosomes. The range primarily reflects the strength of selection against LoF alleles, as well as gene size.

About 15-20% of human genes have such low LoF frequencies that we can infer they are under extremely strong purifying selection; such genes are said to be *mutation intolerant*. Heterozygosity for LoF alleles in such genes must cause substantially decreased reproductive fitness and be associated with serious disorders. These alleles should drive huge increases in risk for these disorders and should be highly enriched for *de novo* events (which can be recognized by family-based sequencing to compare variants in parents and children).

Except for such genes, RVAS in coding regions requires larger sample sizes than CVAS—with at least 25,000 cases needed to have any meaningful power and ideally in the range of 250,000 cases to have good power for most genes. (For roughly two-thirds of human genes, the frequency of people carrying a loss-of-function variant is at least 16 per 250,000 people). Smaller sample sizes may suffice for genes in which LoF alleles are extremely frequent or have huge effect sizes.

RVAS in coding regions: progress to date. The Deciphering Developmental Disorders Project has discovered that severe developmental and intellectual disability are often associated with heterozygous LoF alleles in mutation intolerant genes. As expected, these disorders cause a severe decrease in reproductive fitness and the alleles have huge effect sizes (>100-fold increased risk, approaching but not strictly monogenic). At present, several hundred genes have been implicated at genome-wide significance.

Autism and schizophrenia show significant but much lower contributions from these types of mutations. This is consistent with the extremely high heritability of these phenotypes, in contrast to the low heritability seen for severe intellectual disability. At present, genetic studies of autism and schizophrenia, respectively, have definitively identified roughly 30 and 10 mutation intolerant genes.

For all of these disorders (developmental and intellectual disability, autism and schizophrenia), the vast majority of the variants are de novo mutations—consistent with their having strong effects on phenotypes with strongly reduced reproductive fitness.

Epilepsy and congenital heart disease also demonstrate significant contributions, with some subtypes having more significant contributions than others.

By contrast, cardiometabolic, immune-mediated, and other systemic disorders show no excess of LoF alleles in mutation intolerant genes. In these cases, exome-sequencing studies have identified some genes at the opposite end of the spectrum with respect to mutation tolerance. These studies have found numerous low-frequency common variants with somewhat stronger impact (OR of 1.5-5) that were not captured in early waves of GWAS and imputation. These discoveries have had an outsized impact on our understanding of biology, because they pinpointed specific variants in specific genes. With improvements in variant catalogs, imputation algorithms and biobank scale, these discoveries can and will be made directly from GWAS analysis.

RVAS in non-coding regions. RVAS in non-coding regions is much harder, because there are no clear rules for recognizing functionally important variants or regions over which to aggregate variants. As a result, any true association signals (already requiring very large samples in the coding case) would be greatly diluted, and the required sample size dramatically increased, by one to two orders of magnitude. Moreover, patterns of evolutionary conservation indicate that variants in non-coding regions will typically have much smaller effects than those in coding regions, with the exception of bases affecting splice sites. Unsurprisingly, there has been no progress to date in identifying genome-wide significant non-coding loci by RVAS. Given the challenges in interpreting the tens of thousands of disease-associated *common* variants in non-coding regions, it seems unlikely that contributions from rare variants in non-coding regions will be game-changing.

The best evidence that non-coding regions contain some functionally important rare variants comes from the Deciphering Developmental Disorders (DDD) Project. Whereas 42% of patients with severe developmental disorders carried pathogenic *de novo* mutations in coding regions, analysis suggested that ~2% carried pathogenic *de novo* mutations in highly conserved non-coding regulatory elements. (Analysis suggested ~3% of bases in ~2% of these elements might be pathogenic.)

For RVAS in non-coding regions to ever become practical, huge leaps would be required to precisely identify functionally important variants and regions.

Summary. The emerging picture of common disease genetics is that the majority of heritability is made up of common, largely non-coding variation. However, given our current knowledge of genome biology and the available tools for studying gene function, rare coding variants can make a larger contribution to our biological insight than suggested by the proportion of heritability explained.

2.6. What are we learning: disease biology.

Genetic studies are teaching us a lot about the underlying biological mechanisms of disease.

Disease genes and disease pathways. For well-studied diseases and traits, the lists of hundreds of disease-associated loci (Table 1) have given rise to important biomedical discoveries into the underlying biology of important diseases and traits, including age-related macular degeneration, Alzheimer's disease, cardiovascular disease, inflammatory bowel disease, schizophrenia and other neuropsychiatric traits, and body mass index (see Box 2).

Dissecting different aspects of disease mechanism. By applying GWAS to multiple phenotypes related to a disease, one can learn which loci contribute to which aspects of a disease process. For Type 2 diabetes, about 35% of disease-associated loci show association with reduced insulin secretion in patients. For early heart attack, only 20% of disease-associated loci demonstrate association with lipid levels in patients—indicating that there are many important disease-related processes still to be understood.

Notably, genetic associations with disease susceptibility do not necessarily overlap with genetic associations for disease progression. For example, GWAS studies of lung cancer have identified genetic loci that alter the likelihood that an individual becomes addicted to nicotine—thereby becoming a long-term smoker and increasing life-long mutagen exposure. Perhaps unsurprisingly, this genetic association is not informative for how a patient might progress having already developed lung cancer.

Implicating tissues and cellular processes. A developing set of methods have begun to extend these ideas to explore disease biology, by using the full information provided by GWAS—that is, the genome-wide effect-size vector (defined above). In particular, one might test whether a gene set of interest—such as the genes expressed in a particular tissue or cellular process—is relevant to a disease by studying whether it is correlated with the genome-wide effect-size vector. Such analyses, for example, have clearly implicated the brain in the genetic variation in body-mass index. With the ongoing development of a Human Cell Atlas, it should be possible to extend these studies to implicate not just tissues but individual cell types. This will help connect disease-associated genes to specific cell types and mechanisms active in them.

Exploiting pleiotropy: PheWAS. With GWAS across many different diseases and traits, one can gain insight into the common variants found in one setting to learn about the range of phenotypes associated with the variant. For example, a common variant in *SLC39A8* increases risk for schizophrenia, Crohn's disease and obesity, but decreases risk of hypertension and Parkinson's diseases; the gene encodes a zinc/manganese transporter and a defect in Mn transport may affect protein glycosylation, with different consequences in different tissues. Some loci are associated with a wide range of autoimmune diseases, while others only with specific autoimmune diseases; the former likely play a generalized role in immune balance, while the latter may affect immunity to certain antigens or in certain settings.

By collecting a large cohort with a wide range of phenotypes, researchers can test a set of variants simultaneously against typically thousands of phenotypes; this is referred to as a Phenome-wide Association Study (PheWAS). Notably, PheWAS makes it possible to learn about the range of possible effects of perturbing a gene, which may help anticipate adverse outcomes of a therapy.

2.7. What are we learning: Epidemiology.

Genetic studies are also shedding light on the epidemiology of diseases.

Assessing shared heritability. Beyond studying individual loci, it is becoming clear that much can be learned studying the genome results of GWAS—that is, the genome-wide effect-size vector (defined above). For example, scientists can explore the extent to which diseases involve shared biological processes by examining the correlation between the vectors for the diseases. For example, neuropsychiatric illnesses, schizophrenia, bipolar disorder and major depressive disorder show some shared and some specific genetic effects. These overlaps can be harnessed to boost power and improve polygenic prediction.

Assessing shared environmental exposures and epigenetics effects. Beyond studying shared heritability, the genome-wide effect-size vector can be used to estimate the component of trait variation due to gene-environment interactions, on top of the heritable component of variation. Additionally, evaluating the epigenetic component of heritability, in addition to SNP-level heritability, allows insight into the overall role of gene-environment interactions.

Mendelian randomization. Because genetic variants are fixed at birth, clearly associated variants can be used as instrumental variables to uncover causal relationships across molecular and physiological biomarkers and outcome diseases. For example, Mendelian randomization has shown that while LDL level is a causal risk factor for myocardial infarction, HDL level is not (despite its epidemiological correlation with decreased risk of heart disease).

2.8. What are we learning: Clinical application.

Genetic analysis is also set to have an important impact in the clinic.

Polygenic risk scores. With increasing sample sizes for GWAS, the estimates of the effect size of variants across the genome have improved to the point that in aggregate they can begin to provide clinically useful predictors of common disease susceptibility. These predictors—called polygenic risk scores—can already identify subsets of individuals at substantially higher risk for a number of diseases, including heart disease, obesity, atrial fibrillation, inflammatory bowel disease, and breast cancer. In some cases, the increased risk is comparable to that conferred by Mendelian forms of the disease. Polygenic risk scores will increasingly be used in clinical practice, especially where relevant interventions or screening are available, and to select patients for clinical trials, in order to enrich for events and thereby decrease the required sample size.

For an individual's polygenic risk scores to have maximal accuracy and precision, they should be derived from GWAS data with the appropriate ancestry. As discussed below, a serious issue is that the vast majority of GWAS has been conducted in European-derived populations, with the consequence that polygenic risk scores currently provide poorer predictors for other groups. It is important that this problem is addressed rapidly.

In addition to improving methods to derive scores from CVAS and expanding the representation of ancestries, integrating CVAS risk with RVAS risk will be essential for delivering accurate and complete genetic risk scores. Although rare variants contribute less at a population level, they often contribute a substantial amount for the few individuals that carry a particular mutation. Early data suggest that common and rare variants contribute additively to individual risk in common disease.

Genomically-informed clinical trials. There are important opportunities to improve clinical trials, which are the most expensive element of drug development, by making greater use of genomic information. Genomic information is routinely used in testing cancer drugs, but more rarely used in other areas of medicine. Opportunities include: decreasing the number of patients by using genetic risk information to select patients with a higher background event rate (e.g., rate of heart attacks); increasing the success of clinical trials, by identifying and focusing on subsets of patients in which specific cellular and physiological mechanisms are affected (for example, women and men, or distinct subgroups of type 2 diabetes); decreasing adverse events, by using genomic information (such as through PheWAS and exploratory human trials) to anticipate and understand likely side effects and identify patients at greatest risk; and decreasing time to enrollment, by identifying potential patients through large biobanks with clinical and genomic information.

2.9 Summary: Lessons Learned.

What have we learned so far about the genetics of common disease? Some of the general lessons are summarized in **Box 1**. Examples of insights about specific diseases are given in **Box 2.**

Box 1. Summary of General Lessons

1. Common diseases are highly polygenic. They typically involve hundreds to thousands of common and rare variants across the genome for every disease. Across hundreds of phenotypes studied, the number of genome-wide significant associations with common variants discovered so far exceeds 70,000.

2. These common variants typically have modest effects on disease risk (e.g., altering risk by < 10%) and most lie in non-coding regions (>90%). The non-coding variants primarily act by altering gene expression in one or more cell types (e.g., altering enhancers or splice sites, rather than protein-coding sequences as typical for rare Mendelian diseases).

3. Many common variants influence susceptibility for multiple common diseases. Variants sometimes increase risk for some diseases while decreasing risk for others. Patterns of pleiotropy can shed light on disease mechanisms.

4. Genetic studies of common diseases require extremely large sample sizes.

(i) Common variant association studies (GWAS/CVAS) require many tens of thousands of cases to have reasonable power to detect many loci.

(ii) Rare variant association studies (RVAS) of coding regions require even larger sample sizes—with at least 25,000 cases to have any meaningful power and ideally hundreds of thousands of cases.

For now, well-powered studies will require disease-focused collections. Ultimately, it should be feasible to study many common diseases and traits by combining information from biobanks as total sample sizes climb into the tens of millions.

5. Collectively, common variants likely explain most of the heritability for common diseases. Roughly half of the heritability may be directly due to additive effects of the variants. Genetic interactions among loci may contribute substantial additional heritability, but we currently lack power to detect such interactions. Rare variants and phenotypic heterogeneity (e.g., where two distinct disease entities are combined) may also contribute to explaining the heritability.

6. A genome-wide association study contains a tremendous amount of biological information. We are beginning to learn how to use this information to:

(i) create polygenic risk scores to identify individuals in a population at significantly elevated risk for a common disease (see below concerning the serious underrepresentation of non-European populations);
(ii) assess shared biological mechanisms among diseases; and
(iii) implicate tissues, cell types and biological processes as likely to play a role in a disease.

Box 2. Examples of insights about specific diseases and clinical responses

Age-related Macular Degeneration (AMD). GWAS provided the first mechanistic insight into AMD, showing that genetic variants in genes in the complement system, such as Complement Factor H, play a major role.

Alzheimer's Disease. While research and clinical trials have focused on neurons, GWAS has highlighted the critical importance of microglia, the brain's macrophage cells—leading to a substantial refocusing of drug development efforts onto a cell-type previously considered of little relevance for the disease.

Sickle-cell disease (SCD). GWAS identified the critical role of BCL11A in postnatal expression of fetal hemoglobin, which was already known to substantially modify disease severity—thereby providing a novel therapeutic hypothesis that is being actively pursued.

Inflammatory bowel disease (IBD). GWAS has highlighted the important role of multiple pathways, such as autophagy, which previously had not been implicated in the etiology of IBD. Several of these pathways are now active targets for drug development.

Cardiovascular disease. The epidemiological association between high HDL-cholesterol levels and protection from heart disease had spurred drug companies to invest billions in developing HDL-raising drugs. GWAS indicated that the causative factor was likely not to be HDL, but triglyceride (whose levels are inversely correlated with HDL). Consistent with the human genetics, the clinical trials showed that increasing HDL levels provided no protection.

GWAS studies have also indicated that the majority of the loci that affect heart disease risk do not act by affecting lipid levels—teaching us that there are additional mechanisms to be found. GWAS has enabled effective polygenic risk scores for coronary artery disease, for identifying individuals with substantially elevated risk.

Schizophrenia. GWAS studies have identified an important association with the complement component C4—which plays a key role in the innate immune system, but also is involved in the pruning of synapses in the brain. The result has implicated excessive synaptic pruning during adolescence and early adulthood in the etiology of schizophrenia—providing a target for therapeutic development.

Neuropsychiatric traits and diseases. GWAS studies of diverse neuropsychiatric traits and diseases have identified shared heritability between neurodevelopmental traits and diseases that have highlighted considerable shared genetics among childhood and adult neuropsychiatric diseases - but also unique correlations (for example a strong correlation between the genetics of autism and higher intelligence not seen with other psychiatric traits). These studies have also demonstrated polygenic risk scores that influence the variable expressivity in monogenic rare neurodevelopmental disorders.

Body Mass Index. GWAS studies have provided strong support for a role of the central nervous system in susceptibility to obesity. A particularly interesting example is the melanocortin-4 receptor (MC4R pathway), in which both non-coding and coding variants are associated with variation in BMI across the population.

Atrial fibrillation. Whereas atrial fibrillation has long been thought to be due to dysfunction of ion channels, genetic studies have identified other potential mechanisms for the arrhythmia. Specifically, CVAS and RVAS have both identified variants in or connected to genes encoding parts of the sarcomere or contractile apparatus of the cardiomyocyte. These studies suggest that at least some forms of atrial fibrillation may reflect a subtle atrial cardiomyopathy rather than a channelopathy. This new finding may have clinical significance and is an area of active investigation to determine if these patients have a differential response to antiarrhythmic medications, procedures or the need for anticoagulation.

Gene x Environment interactions. GxE interactions are known for various common complex diseases. Studies have shown an increased risk of Parkinson's disease among individuals who are exposed to various pesticides and who are slow metabolizers for CYP2D6 or PON1. Genetic variation at NOS2 has been reported to interact with residential traffic-related exposure to affect exhaled nitric oxide (FeNO) levels, which are a known biomarker of airway inflammation. Genetic variation at NAT2 (which catalyzes metabolism of aromatic monoamines) increases risk of bladder cancer among smokers but not among never-smokers. Genetic variation in ALDH2 (which detoxifies acetaldehyde) interacts with alcohol exposure to increase the risk of esophageal squamous cell carcinoma.

3. Understanding the Human Genome

Understanding common diseases requires more than genetic mapping. An important first step involves connecting disease-associated loci to their immediate targets in the human genome—i.e., identifying the causal variants at the loci, the functional elements they alter, and the genes that directly affect in the relevant cell types.

Progress in understanding common disease genetics has thus depended on rich knowledge about the structure and function of the human genome—resulting from comprehensive projects undertaken by the scientific community.

By 2003, the Human Genome Project had produced a high-quality genome assembly containing the vast majority of the euchromatic sequence of the human chromosomes and a parallel effort, the SNP Consortium, had produced a catalog of 1.4 million common genetic variants.

But, there was still much to do to annotate the human genome—including (i) improving the genome sequence itself, (ii) expanding the catalog of genetic variants, (iii) refining the gene catalogue, and (iv) characterizing how the genome is read in different contexts, including the patterns of gene expression, active regulatory elements and chromatin structure. The latter is crucial for connecting variants to their immediate function.

A wide range of projects have played—and will continue to play—a crucial role in filling out this knowledge. We outline some of these efforts below.

3.1 Human Reference Sequence.

Efforts have continued to refine the human reference sequence, including filling in highly repetitive sequences that have been hard to sequence and assemble.

The current human reference genome (GRCh38) is the most accurate and complete vertebrate genome ever produced, after iterations of filling in the missing gaps (ribosomal rDNA arrays, large segmental duplications, satellite DNA arrays etc.). There remains great interest in closing the remaining gaps, which can lead to experimental artefacts and harbour unexplored variation. Current long-range sequencing methods hold great promise in this regard.

***3.2 Catalogs of Genetic Variation*.**

The successful mapping of common diseases and traits requires near-complete knowledge of common genetic variation. Several large-scale international collaborations—the International HapMap Consortium, the 1,000 Genomes Project, the Haplotype Reference Consortium (HRC), the Exome Aggregation Consortium (ExAC) and the Genome Aggregation Database (gnomAD)—have played central roles in generating, analyzing and disseminating reference catalogs of variants, their frequency, and their patterns of linkage disequilibrium patterns, and have made it possible to use genotypes at a subset of genetic variants (e.g., 500,000 ‘tag SNPs’) to impute most of the rest.

These publicly available resources now include a catalog of ~20 million SNPs and small indels. They include >99.9% of all SNPs with frequency ≥1% in European-ancestry populations, but a lower fraction for other populations and especially for African populations (see below). Various methods and software have been developed to perform imputation, such as IMPUTE, MaCH, Beagle and others, and have become a standard part of current CVAS.

The resources (especially, ExAC and gnomAD) are widely used by clinical geneticists in their quest to identify causal variants in families with rare diseases, by excluding variants that are too frequent.

They are also being used to infer the degree of purifying selection on each gene based on the number and frequency spectrum of variants. With increasing sample size and population diversity, it is becoming possible even to infer selective constraints on individual exons and smaller segments encoding portions of protein.

Much still remains to be done.

It will be important to saturate SNPs at lower frequencies, to improve genetic imputation, genetic understanding (such variants tend to have somewhat larger effect sizes) and clinical genetic interpretation. Efforts are also needed to achieve more complete coverage of other types of variants, including large deletions, duplications and inversions.

Most importantly, we need to increase the number of individuals sequenced from non-European-ancestry populations to enable variant discovery, imputation and design of genotyping arrays at comparable levels as for European-ancestry populations. Increasing the study of African populations is especially important. Clinically, African populations have been underserved. Genetically, African populations have so much to teach us because Africa has greater genetic diversity and shorter haplotypes (providing better mapping resolution) than other continental populations.

3.3. Catalogues of gene expression for tissues and cell types.

A key challenge, as noted above, is to understand how disease-associated variants directly affect cellular function.

One important foundation would be to know the expression patterns of all genes across all cell types (including transcript levels and splicing) and how these expression patterns are altered by *cis*-acting genetic variants (that is, variants carried on the same physical chromosome). The latter is important because most disease-associated loci lie in non-coding regions and are likely to act by altering the expression of nearby genes.

To help assemble this knowledge, the genotype-tissue expression (GTEx) project was launched. The project has collected samples from 54 tissue sites across nearly 1,000 individuals and has performed molecular assays including genotyping and imputation, genome sequencing, and RNA sequencing.

Beyond creating a catalog of 'typical' gene expression in the bulk tissue samples, the GTEx Project has applied GWAS to associate inter-individual variation in the RNA expression of genes with nearby genetic variants — identifying *cis*-expression quantitative trait loci (*cis*-eQTLs). Notably, a subset of these cis-eQTLs have tissue-specific effects on gene expression. cis-eQTLs have been identified for nearly all coding regions. These *cis*-eQTLs may be good candidates to explain many disease-associated loci. Similarly, RNA expression catalogs are being used to map variants that affect splicing of particular transcripts, called splice QTLs (sQTLs).

Much important information is lost in the heterogeneity of bulk tissues. In the years ahead, it will be important to extend the work to the level of single-cell analysis—using insights and methods being developed by the Human Cell Atlas.

3.4 Epigenomic catalogues for tissues and cell types.

A second important foundation would be to know all active regulatory elements in all cell types, because many non-coding variants are likely to act by affecting these elements. Studies of disease-associated loci often focus on tissues and cell types in which the locus lies in an active regulatory region (although this approach is limited because current coverage of tissues and cell types remains incomplete). Notably, the functional elements identified so far are enriched in GWAS signals.

The Encyclopedia of DNA Elements (ENCODE) Project was the first effort to systematically discover the functional elements (both coding and non-coding) in the human genome. Comparative genomic studies have suggested that ~8% of the human genome is under purifying selection and thus likely functional. ENCODE analyzed 147 different cell lines/cell types with systematic assays for transcription, transcription factor binding, histone modification and chromatin accessibility. The project released 1640 data sets in 2012, and the data have continued to grow. By providing an initial map of regions with transcription, promoters, putative enhancers across many cell lines and cell types, ENCODE has given important clues for understanding disease-associated loci.

The Roadmap Epigenomics Mapping Consortium has expanded on this work, by investigating stem cells and primary *ex vivo* tissues selected to represent the normal counterparts of tissues and organ systems frequently involved in human disease. Using various sequencing approaches, the consortium has characterized DNA methylation, histone modifications, chromatin accessibility and short and long non-coding RNA transcripts. Integrated analysis has provided high-resolution maps of putative regulatory elements spanning the cells and tissues—highlighting epigenomic differences across lineage and differentiation and relationships between enhancers, promoters and transcripts.

As with RNA expression, much of the important epigenomic signal is obscured by the heterogeneity of bulk tissues. It will be important to extend this work to the single-cell level. Some epigenomic methods have been adapted to single-cell analysis (for example, ATAC-seq), but considerable work is still needed for others. Moreover, it will be important to undertake these studies at a population level, to capture human variation and diversity.

The impacts of environmental exposures may also be important in understanding epigenomic and transcriptomic patterns. Efforts to understand exogenous environmental impacts on regulatory pathways, including loci that promote susceptibility or resistance to exposure risks, may inform the context in which these processes affect disease prevention and treatment.

To link common diseases to cell-specific mechanisms, we will ultimately need a clear and comprehensive picture of genome regulation at the single-cell level across all cell types in large numbers of individuals.

4. Disease Mechanisms and Clinical Translation

While much can be learned from “generic” information about the genome, transcriptome and epigenomic, deciphering disease mechanisms and developing therapies also requires an intense focus on disease-specific information. While the details will differ, there are still great opportunities for shared learning across diseases about the most effective paradigms.

4.1 Disease Mechanisms.

It is becoming clear that understanding disease mechanisms will benefit from:

- gathering genetic data for many phenotypes of translational interest, including those related to disease progression, complication risk, therapeutic response and clinical outcome.
- gathering large-scale data on the cell-types and cell states most relevant to the disease. Generic approaches will not characterize all cellular contexts relevant to a disease (e.g., exposure to secretagogues, immune activation, pathogens, toxins) deemed relevant to specific diseases.
- interpreting large-scale functional data in light of known disease biology. While genome-wide analyses can implicate particular tissues, cell types, and cell processes, these enrichments do not mean that all of the physiology of a disease (and all of the GWAS loci) will be mediated through these. Disease-specific knowledge will provide an important filter for interpreting this information and ensuring that functional data can be generated and interpreted in ways that are appropriate for each disease, and at each locus for that disease.
- developing precise cellular assays, to assign genes to disease-specific processes. For example, one might characterize genes in IBD-associated loci by applying CRISPR-based screens to identify all genes whose perturbation affects disease-related processes, such as autophagy.

4.2 Clinical translation.

There is growing interest in the pharmaceutical industry in using human genetics to validate targets (**Box 3**). Various types of knowledge are valuable for supporting clinical translation.

Mechanisms. Understanding the mechanism of action for as many loci as possible for each disease is key to understanding disease pathology. The aim should be to pursue such efforts until the point that the biology is “saturated”—in the sense that most additional loci correspond to already-encountered pathways. Various high-throughput functional assays should be developed to explore, manipulate, and monitor the pathways following genetic or therapeutic perturbation *in vitro* and *in vivo*. These assays are critical in drug development for supporting molecular screens, lead-molecule optimization, and enabling more detailed mechanistic investigation.

Direction and magnitude. It is important to have information about a wide allelic spectrum to ascertain the desired direction of therapeutic effect (does lower activity of a target confer lower or higher disease risk?). While common variants with modest effects can point to excellent therapeutic targets (e.g., the targets of statins and sulphonylureas), it is valuable to identify (often rarer) human alleles with larger effects—to guide the extent of therapeutic modulation that would be required for clinical benefit and to provide clues about a useful therapeutic window.

Biomarkers. Genetics can help to identify translational biomarkers, linked to the molecular mechanism of interest. Such biomarkers are an essential component for successful drug development, as well as for supporting clinical and epidemiological analyses.

Pharmacodynamic biomarkers can be particularly important as readouts of target engagement (providing a quantitative assessment of the magnitude of therapeutic perturbation achieved) in early, proof-of-concept clinical trials. Biomarkers of clinical efficacy can provide reliable intermediate surrogates for clinical end-points that are difficult or costly to capture in prospective trials. Other biomarkers can be used to stratify patients on the basis of high future risk of disease or particular disease subtype, or to signal toxicity

Therapeutic modality. Information on disease mechanism, tissue of action, and the characteristics of a candidate therapeutic targets can provide information about potential “druggability” and highlight the most appropriate therapeutic modality (small molecules, biologics, or other innovative approaches).

Pleiotropy. As noted above, many disease-associated variants influence traits other than those related to the index disease. For therapeutic targets, this information may point to potential toxicities or adverse events—or, in some cases, opportunities for drug repurposing. (Of course, the pleiotropic spectrum of the often-tissue-specific variants typically captured by GWAS may be more restricted than effects observed when the same target is perturbed across all the tissues.)

Patient stratification. Matching drugs to their optimal target population requires deeper understanding of individual risk (of disease onset and disease progression), as well as of etiological and clinical heterogeneity. Such knowledge is emerging from the analysis of large-scale cohort studies. A challenge is to combine genetic, biomarker, lifestyle, environmental exposure and clinical data to best stratify risk and disease subtype in ways that may have clinical utility (“precision medicine”). For example, the capacity to stratify patients can be critical for Phase 2 studies (e.g. selecting those with the highest background risk of disease progression), and then for defining patients most likely to benefit from an approved medication (in terms of maximizing efficacy, minimizing adverse events, and/or addressing unmet clinical needs).

Box 3. Human genetics and drug development

Drug development faces two major challenges. First, the cost of delivering a successful new medicine to the market is high—estimated at >$2B US per commercial launch, with costs having increased by two-thirds from 2010 to 2017 (Ref 1). A major contributor is attrition—the fact that so many prospective medicines fail in clinical development. Second, many new medicines fail to make major improvements above the standard of care.

Human genetics holds potential to address these problems, as supported by several lines of evidence (Ref. 2):

- There is anecdotal evidence that human genetics can drive drug discovery and development. A good example is PCSK9, a target identified by human genetics for which there are two approved therapies that block circulating PCSK9 protein levels, lower the levels of circulating LDL cholesterol, and protect from cardiovascular disease.
- There is support from success rates at individual drug companies. In 2014, Cook et al. published the results of a comprehensive longitudinal review of AstraZeneca's small-molecule drug projects from 2005 to 2010 (Ref. 3). They found that therapies against targets with human genetics were roughly twice as likely to lead to successful programs compared to therapies against targets without human genetics support.
- There is support at the level of all approved medicines. In 2015, Nelson et al. estimated that selecting genetically supported targets could double the success rate in clinical development (Ref. 4). More recently, King et al extended these findings (Ref. 5). Both studies demonstrated that probability of success was greatest when a molecular mechanism of action could be quantitatively matched to a therapeutic modality specific to the clinical indication of interest.
- There is support at the level of indications. The success rate from Phase 1 to approved medicines are more than twice as high for rare genetic diseases compared to chronic, high prevalent diseases (25% vs 9%).

These factors have led the life sciences industry to increasingly incorporate human genetic evidence as a key pillar in R&D strategies. As a consequence, the proportion of pipelines with human genetic evidence is steadily increasing. For example, Amgen states 75% of their portfolio is grounded in human genetics, a number that is consistent with a number of other industry partners involved in the ICDA. There has also been an increase in the number of companies pursuing multiple indications in parallel for genetically well-defined targets. A striking example is TYK2: Pfizer is currently running five parallel Phase 2 clinical trials (psoriasis, ulcerative colitis, Crohn’s disease, vitiligo, and alopecia areata) for a TYK2/JAK1 inhibitor, and Bristol-Myers Squibb is running five Phase 2/3 clinical trials (psoriasis, psoriatic arthritis, ulcerative colitis, Crohn’s disease, systemic lupus erythematosus) for a selective TYK2 inhibitor. By contrast, traditional clinical development programs would be more risk averse, testing new indications sequentially rather than in parallel.

1. Deloitte: “Embracing the future of work to unlock R&D productivity” https://www2.deloitte.com/uk/en/pages/life-sciences-and-healthcare/articles/measuring-return-from-pharmaceutical-innovation.html
2. Plengegen.com blog – need to write!
3. Cook et al (2014) Nature Reviews Drug Discovery (https://www.nature.com/articles/nrd4309)
4. Nelson et al (2015) Nature Genetics (https://www.nature.com/articles/ng.3314)
5. King et al (2019) biorxiv (https://www.biorxiv.org/content/10.1101/513945v1)

5. The Goal Ahead: From Maps to Mechanisms to Medicine

Common disease genetics now stands at a pivotal moment. The field has a reliable paradigm to map loci containing variants affecting any common disease or trait, which has so far revealed more than 70,000 loci related to hundreds of diseases and traits.

However, a huge challenge remains: It remains far too difficult to move from disease-associated loci, to disease biology, to disease treatment. Even the first step has been accomplished only in a small fraction of cases to date.

To accelerate progress, we need to solve the "*Maps to Mechanisms to Medicine*" Challenge.

Why is this the right time? Powerful scientific driving forces—in human genetics, cell biology, and data science—make this the right time to tackle this challenge (**Box 4**).

Box 4. Scientific driving forces enabling the next phase of human genetics

Human genetics is being propelled by:

- growing efforts around the world to create large biobanks;
- increasing deployment of genetics in medical practice;
- continuing decreases in the cost of genome sequencing, likely to fall to around $100 per genome in the next several years; and
- improvements in statistical methods for analysing genomic data.

Cell Biology is being propelled by:

- the revolution in single-cell analysis (including genome-wide assays of RNA expression and chromatin features in isolated cells and, increasingly, in histological samples), propelling a Human Cell Atlas project that will yield a comprehensive characterization of all human cell types;
- the revolution in genome editing, using CRISPR systems, that has made it possible to directly test the effects of specific genetic variants and specific gene disruptions in human cells; and
- new biological models for studying human cell types, including induced pluripotent cells and organoids.

Data Science is being propelled by:

- the revolution in machine learning, which has made it possible to use massive datasets to make effective predictions in many fields and, of special importance for biology, might be adapted to reveal underlying causal mechanisms;
- the availability of cloud-native data platforms, enabling scientific communities to store and analyze massive datasets in centralized settings;
- active development of technologies for federated data analysis by multiple parties, designed to provide security and respect privacy;
- the development of robust standards for genomic, phenotypic and clinical data; and a

prevailing ethos of data sharing.

To move from Maps to Mechanisms to Medicine (M2M2M), we need approaches to address a wide range of issues. Notably, the issues cannot be addressed in a strictly linear order. In many cases, the solution to one will likely be informed by progress on the others. For example, knowing the relevant cell types for a disease will help inform the identification of the relevant genes, and *vice versa*.

The scientific agenda is summarized in Box 5.

Box 5. Maps to Mechanisms to Medicines (M2M2M) Challenge: Overview

To move rapidly from the discovery of loci harboring disease-associated variants to understand disease mechanisms and propel therapeutic development, we must develop systematic ways to discover:

- the causal variants at loci that affect disease susceptibility, disease progression, and therapeutic responses;
- the immediate molecular effect of these variants (e.g., whether they act by altering a binding site for a transcription factor in an enhancer, a splice site or a protein sequence);
- the target genes on which these variants act;
- the cell types and states in which the causal genes operate;
- the cellular processes and physiology through which the genes act along the pathway to disease;
- a wide spectrum of causal alleles—from weak to complete loss of function, as well as gain of function when present;
- the effects of the variants on other phenotypes, both related and unrelated to the disease;
- effective cellular and animal models to aid study of disease processes;
- biomarkers of the disease process and progression;
- functional assays to support therapeutic development; and
- predictors to identify individuals at high risk of disease incidence or progression, applicable to all relevant populations—to improve research, increase efficiency and innovation in clinical trials (by incorporating genomics), and most importantly, improve care.

We must also develop:

- data platforms that allow the community to store and analyze data in cloud-based systems accessible to any authorized investigator;
- methods to enable federated analyses, making it possible to learn from diverse datasets without compromising privacy or security; and
- high standards for clinical services, ensuring that best practices for interpreting genetic

information are available to all providers and patients.

As with the Human Genome Project, which enabled the discovery of the genes responsible for rare Mendelian diseases, and the GWAS revolution, which enabled the discovery of loci underlying common diseases, solving the M2M2M Challenge will involve the international human genetics community coming together under a shared vision to develop new ideas and principles, undertake foundational research projects, create comprehensive catalogs, develop new technologies and democratize these resources.

Below, we elaborate on the components of the M2M2M Challenge:

Reliable methods to discover the causal variants and their immediate molecular effect. Disease associations identify haplotypes that often contain tens of variants in strong linkage disequilibrium across many kilobases. Pinpointing which variant(s) in the locus plays a causal role is difficult. In the less than 10% of cases where the disease-associated variants include a coding variant, one has a reasonable guess but no guarantee. In the vast majority of cases where the variants lie entirely in non-coding regions, the challenge is especially hard due to our limited ability to recognize functional elements and identify causal variants within them.

One approach is to perform “fine mapping” to distinguish between variants in close linkage disequilibrium, by substantially increasing the sample size in the initial population studied. Combining information from diverse populations with differing alleles and haplotype structures can be particularly powerful in fine mapping. African populations may be particularly helpful in this regard, because haplotypes tend to be shorter.

In a particularly favorable scenario (involving analysis of more than 30,000 IBD cases), high-resolution fine-mapping identified a single SNP that was more than 50% likely to be the causal variant for roughly one-third of the genome-wide significant loci. Still, in the clear majority of cases, fine mapping will not narrow loci to single variants.

Functional information will be important in identifying a causal variant(s). Ultimately, we want to have a comprehensive catalog of *all* functional elements in *all* cell types and states (such as enhancers, promoters, splice sites, CTCF sites, etc.), together with reliable experimental and/or computational ways to assess whether and how a variant significantly alters function (e.g., altering a transcription factor’s binding site, chromatin structure, protein function, etc.)

Reliable methods to identify the target gene(s) affected by a causal variant. A crucial issue is to identify the target gene(s) through which a variant modulates the disease mechanism, including the magnitude and direction of effect. This task is complicated by the fact that a regulatory variant may affect many genes, of which only a subset may be causally related to disease. In some cases, key information may come from variants present only in certain ancestry groups, underscoring the need to study non-European populations.

Reliable methods to discover key cell types and states through which the causal variants and genes act. Knowing the relevant cell type to study is critical for discovering and characterizing the role of causal variants, understanding disease mechanisms, and developing therapeutics.

The Human Cell Atlas project is working to develop a comprehensive picture of all human cell types (including throughout development), with respect to gene expression, aspects of chromatin state, and three-dimensional organization in tissues. This information is important for identifying and studying individual loci—for example, to know *all* cell types in which a given enhancer is active. It will also be essential for inferring relevant cell types from ‘bulk’ signals, particularly the genome-wide effect-size vector. The level of resolution that ultimately can be obtained (individual cell types vs. sets of related cell types) remains to be seen.

Considerable care and thoroughness will be required. Some variants may act not only in specific cell *types*, but only in specific cell *states* (such as homeostatic states, developmental contexts, and exposures) or in establishing cell-type proportions. Importantly, cell types causally involved in disease etiology may differ from those responsible for disease symptoms; understanding the difference will be crucial for targeting interventions to root causes rather than symptoms.

Reliable methods to discover the target cellular programs through which the causal variants and genes act. Understanding the “cellular programs” affected by the disease-associated variants and their target genes is fundamental for understanding disease etiology and pathophysiology. (An important aspect of this work will be defining precisely what should be meant by cellular programs.) It will also be key to developing and targeting interventions and for understanding possible biological redundancy and compensatory mechanisms that may affect the outcome of pharmacological intervention.

In several well-studied common diseases, a subset of the disease-associated variants and genes clearly converge on certain cellular programs (for example, multiple GWAS hits in IBD related to autophagy, and several in schizophrenia related to synaptic pruning). However, the generality of this observation remains to be established.

We lack well-established methods for tracing causal cellular programs. One approach is to modulate target genes, using CRISPR-based approaches, in relevant cellular, organoid and animal model systems and then monitor the transcriptome-wide effects on gene expression. A complementary approach is to learn from human patients, by studying gene-expression patterns in pre-symptomatic individuals with high polygenic risk scores. (One cannot rely on gene-expression patterns in symptomatic individuals, as most of these changes are likely to be consequences rather than causes of the disease.)

These studies would ultimately be aided by the creation of a “Comprehensive Catalog of Cellular Programs”, which might be created based on information from the ongoing Human Cell Atlas project.

Rich allelic spectrum of disease-causing variants for the disease. Genetic studies in all organisms benefit from having the ability to study a rich spectrum of alleles—from weak effects, strong effects and complete loss of function. Large GWAS studies, primarily in European-ancestry populations, have highlighted many common variants with modest effects. However, many steps need to be taken to fill in the allelic spectrum. We need:

- *continued studies in European-ancestry populations*, to identify more of the loci that are currently below statistical significance.
- *comparably large studies in non-European-ancestry populations*, which will yield disease- and trait-associated loci that has been missed in European-ancestry studies

due to differences in allele frequencies (owing to both genetic drift and population-specific selection pressures). For example, a GWAS for type 2 diabetes in Latino-ancestry individuals found that the strongest effect in the genome was at *SLC16A11*, a gene that was not detected in previous European-ancestry studies, where the allele frequency of the variants is 25-fold lower. Similarly, a GWAS in Latinas discovered an ESR1 allele that arose in indigenous Americans and confers a strong protective effect against breast cancer. Furthermore, some important phenotypes can only be studied in populations with certain endemic exposures—for example, susceptibility and resistance to certain infectious diseases. (As noted above, additional information from European-ancestry and non-European-ancestry populations will also aid in fine-structure mapping and potentially in identifying allelic series.)

- *studies in founder populations,* where a population bottleneck will allow many deleterious alleles to reach high frequencies. For example, the current Finnish population contains thousands of deleterious coding variants at high frequency (>1%) — which can be readily discovered and studied.
- *studies in populations with high rates of consanguineous marriages,* where it is possible to study recessive effects of low-frequency deleterious alleles. Ideally, we would like to understand the effect of homozygous loss-of-function (‘knock-out phenotype’) for all human genes.
- *large-scale discovery of rare variants,* to identify strong-effect alleles. As noted above, well-powered RVAS for coding regions will require analysis of hundreds of thousands of patients for most common diseases. Studying diverse groups is likely to increase the probability and expand the range of discovery.

Overcoming population biases. The fact that the vast majority of GWAS has occurred in populations with predominantly European ancestry has implications not only for gene discovery, but for clinical use. Specifically, the predictive power of polygenic risk scores in underrepresented populations is substantially lower — which may exacerbate health inequalities as these scores are integrated into clinical practice. Moreover, as polygenic risk scores become increasingly used in functional studies, there is a risk that these population biases become further entrenched in biological studies.

Identifying the full range of effects of a variant. Understanding the full range of effects of a causal variant can shed light on the underlying pathophysiology, as well as informing the safety issues associated with drugging the target. We will need to understand the effects of variants on all aspects of a disease (susceptibility, progression and response to clinical interventions) and also on many other diseases and traits. There is growing evidence for extensive gene-environment interactions, which can illuminate disease biology and suggest how modification of exposures and lifestyle factors might reduce disease risk.

Ultimately, we will want to carry out PheWAS for all medically relevant phenotypes and millions of people across many medical systems. At present, though, PheWAS is laborious and incomplete— largely due to difficulties in accessing and cross-analyzing data, and lack of data for many phenotypes.

Effective disease models. We will need to create mechanistically-accurate models of disease, using human tissue, human organoids created in vitro and animal models in vivo, as well as functional assays to interrogate these models for processes relevant to disease.

Biomarkers. As described above, we need ways to use genetics to identify clinical biomarkers connected to underlying mechanisms to facilitate monitoring of drug efficacy and selection of patients.

Powerful and widely accessible data platforms. We need high-quality, widely accessible data platforms that enable cloud-native data storage of human genetic data and analysis with best-practices pipelines and analytical tools.

Methods to combine information from diverse sources, while maintaining security and patient privacy. There are various potential approaches, including methods based on policy (such as restricting access together with laws or contracts forbidding deidentification) computer science-based methods (such as secure multiparty computation and homomorphic encryption) that provide mathematical guarantees in specific circumstances.

High standards for clinical services. Finally, the human genetics community has a responsibility to ensure that high-quality, clear information is readily available for use in clinical services, and that genetic information delivered via clinical services is high quality.

What would success look like? The ambitious goal of the M2M2M era would be:

- The discovery of the key mechanisms underlying the etiology and progression of most common diseases.
- A revolution in drug development programs for common diseases, built upon the knowledge of causal mechanisms and informed by genetic evidence about the safety and efficacy *in vivo*.
- A transformation in clinical care for many common diseases by using genetic information to determine which therapeutic options are most likely to be efficacious and safe for individual patients.
- The ability of individuals to know the common diseases to which they are predisposed, and to make lifestyle and other choices to maximize their health and well-being.

Such progress will not be simple or rapid: As with other phases in human genetics, realizing the promise may require 15 years. But the lasting impact will be tremendous.

From a **researcher's perspective**, they will see a transformation in the ability to answer questions. Analyses that were previously impossible will become possible. Analyses that today are slow, costly and laborious will become scalable, robust and standardized. Analyses that required generating experimental data will become rapid *in silico* 'look-ups'. Researchers will begin to be able to read the functional code of the human genome and understand how it underlies human biology and diversity.

From the **patient's perspective**, she could get earlier information on which diseases she is likely to suffer from in later life, and make choices that lessen her risk. If being treated for a common disease, she could be more confident that she will receive a drug that is more likely to benefit her and less likely to harm her. She would also have more opportunities to participate in research relevant to her condition.

From a **doctor's perspective**, she will have better drugs with which to treat her patients, can be more confident that the drugs she prescribes are more likely to provide a net benefit to her patients, and she can advise them more accurately on their likely prognosis.

From a **drug company's perspective**, it will have hundreds of genetically-validated drug targets, associated with information about disease-relevant cell-type(s) and genetic 'safety profiles' pointing to likely side effects. For clinical trials, it will be able to increase efficiency by using predictive genetic stratification to select patients and molecular biomarkers of disease progression to monitor response and provide confidence in presumed mechanisms of action.

From a **healthcare payer's perspective**, it will be able to deploy more efficacious drugs more cost effectively. Moreover, it will be able to identify individuals at high risk of disease earlier and intervene earlier, prolonging healthy life.

6. International Common Disease Alliance

Human genetics has a history of quantum leaps of productivity, spurred by deliberate efforts by the human genetics community to articulate a shared, ambitious vision. Such efforts in the mid-1980s led to the Human Genome Project and in the late 1990s and early 2000s to the GWAS revolution.

Once again, there is a growing sense across the human genetics community that it is the right time to articulate a vision for the next phase of common disease genetics. Over the past two years, discussions among scientists across the human genetics community have led to the decision to form an International Common Disease Alliance (ICDA) as a way to engage the community. ICDA was officially launched at a September 2019 meeting near Washington, DC.

Role of ICDA. The International Common Disease Alliance serves as a **scientific forum** bringing together international stakeholders across academia, medicine, biopharma companies, tech companies, and biomedical funders to:

- **define current barriers to progress in tackling the M2M2M challenge,** including scientific, technological, policy, computational and organizational obstacles;
- **identify needs and opportunities for new projects** to overcome these barriers in the spirit of past and present examples of public efforts and public-private projects (such as the SNP consortium, the HapMap projects, the 1000 genomes projects, the FinnGen Project, the UK BioBank, the All of Us Project, the Open Targets Initiative, and many more).
- **organize working groups to propose solutions and to drive progress**, including key knowledge, datasets, experimental technologies, computational platforms, and frameworks for data sharing and data harmonization;
- **organize scientific meetings to bring together the community** on an ongoing basis to share results, assess progress, and update plans about the genetics of common disease;
- **coordinate with funders** to ensure the work defining the barriers and proposed solutions, in the white papers are of maximal utility;
- **help to facilitate international collaborations**, where appropriate; and
- **undertake public communication and engagement** on issues related to common disease genetics.

ICDA itself does not expect to directly decide upon or fund scientific projects. Rather, we anticipate that the scientific members of ICDA will undertake collaborative initiatives, such as piloting new experimental technologies to tackle critical challenges, scaling up promising technologies to generate foundational genomic data resources, developing novel analytical methods to integrate large genetic and genomic datasets, developing new data platforms, and developing ethical frameworks for truly global collaboration.

Role of partnerships. Partnerships of many kinds will be critical for success, involving clinicians and researchers, epidemiologists and geneticists, scientists who study environmental exposure, statisticians and wet-lab biologists, technology developers and data generators, academic and government research institutions, hospitals and other medical institutions, biopharma and tech companies, national and philanthropic funders, and scientific journals.

ICDA will work in partnership with the human genetics community; not seeking to duplicate the many highly functional activities already underway, but rather serving as a scientific venue for discussing, stimulating, and sometimes coordinating activities.

Chapter 2. Scientific Goals and Foundational Resources

1. Looking Ahead

The **ICDA White Paper** is intended to be a living document, which will evolve based on input from ICDA Working Groups and the broader community. The first chapter aimed to set the stage by reviewing the history and describing the important challenges and opportunities ahead. The remainder of the white paper aims to outline a clear vision for the future, including concrete proposals for how to achieve it.

The ICDA community aims to define four things:

(i) **Foundational scientific knowledge** that would dramatically accelerate the understanding and treatment of common diseases. Identifying the scientific questions to focus on involves balancing two issues: What would be most transformative, and what might be feasible?

(ii) **Foundational scientific resources** that we might create to drive progress, including toward creating the foundational scientific knowledge. By foundational resources, we include clear scientific concepts, comprehensive and sustainable datasets, experimental technologies, analytical methods, computational tools, and data platforms. As described in the history above, wise choices of projects to create foundational resources have played a critical role in driving progress in human genetic and genomic research over the past 35 years.

(iii) **Key disease applications** that would be well suited to pioneer the development of solutions.

(iv) **Clear recommendations** for how these goals might best be accomplished.

This White Paper makes a start at the first three items above. It was developed (in September 2019, updated in February 2020, and v1.0 finalized in May 2020) with the aim of beginning discussion within the ICDA community about specific plans that would best address the community's needs.

In May 2020, ICDA released Recommendations (v1.0) to the scientific community.

2. Overarching Scientific Goals

To propel the understanding and treatment of common diseases, we must be able to move rapidly from Maps to Mechanisms to Medicine (M2M2M Challenge). At a high level, we would ideally like to have certain knowledge at our fingertips:

Goal 1: Know the phenotypic consequences of any variant in the human genome.

It would be tremendously valuable to know the phenotypic contribution of *any* variant in the human genome — whether it is a common variant in the human population, a rare variant found only infrequently, or a variant that has not yet been observed. (We note that all non-lethal single-nucleotide variants likely exist in the human population, given the number of genomes in the human population (~1.5 x 10^{10}) and mutation rate per nucleotide per generation (~1.3 x 10^{-8})—representing an extraordinary trove of biological information.)

If interpreted literally, the goal is surely not feasible in the foreseeable future. However, tremendous progress can be made toward the goal.

For common variants, we could get very far toward the goal based on direct empirical observation of the human population — creating the 'genotype x phenotype' matrix across tens to hundreds of millions of people. The greatest challenge is no longer obtaining genetic information: whole-genome sequencing will soon fall below $100, a tiny fraction of lifetime healthcare costs for individuals in most countries (although not all). The greater challenges will be (i) ensuring the ability and utility of gathering and integrating genetic and medical information for individual patients, (ii) enabling federated analyses that protect the security and privacy of patients' data, and (iii) earning trust among participants.

For rare variants and never-before-seen variants, the answers will surely be more speculative. They will require ways to combine (i) empirical data on similar variants with (ii) extensive functional knowledge of the human genome.

Goal 2. Know the functional elements encoded in the human genome, and understand their functional constraints.

It would be tremendously valuable to know all functional elements (including protein-coding transcripts, functionally important non-coding transcripts, splice sites, promoters, enhancers, CTCF sites and any other major classes), as well as all cell types in which they are active and the chromatin state at those elements.

Over the past 15 years, there has been important progress toward this goal — especially for individual cell lines and human tissues (through such projects as ENCODE, Epigenetic Roadmap, and GTEx). With recent advances (including single-cell biology) and decreasing costs, it is becoming feasible to create a high-quality catalog for the human body.

The greater challenge will be to understand the functional constraints on these elements — that is, which nucleotides play which roles. Because it will not be possible to mutate and assay every nucleotide, gaining this knowledge will require multiple approaches to infer constraints—including drawing on natural variation, experimental perturbations, evolutionary conservation, machine learning, and more. The goal is not just to know the constraints, but to understand the reason for them—for example, to know precisely which transcription factors are binding at each enhancer in each cellular context.

Goal 3. Know all human cell types and all cellular programs, in health and disease.

It would be tremendously valuable to know all cell types in the human body, as well as the cellular trajectories that lead to the cell types and the various cell states and contexts in which the cell types occur. The Human Cell Atlas project is characterizing and cataloging cells in healthy human tissues by using various single-cell molecular signatures (including single-cell transcriptomics and chromatin analysis) and is increasingly using *in situ* transcriptomics to understand spatial relationships among cell types. In addition to healthy tissues, it will be important to have comparable atlases of relevant tissues in the setting of diseases.

An even greater challenge will be to use rich single-cell data to systematically infer all 'cellular programs.' By cellular programs, we loosely mean the circuitry that propels cells to develop into particular cell types, shift to particular cell states or remain stable. At least at the level of gene regulation, we should aim to be able to comprehensively recognize all cellular programs and understand the role of chromatin architecture and accessibility, specific transcription factors, regulatory elements, and target genes in these processes. Such a comprehensive view would greatly assist in connecting disease genetics to disease biology.

Goal 4. Know the processes that mediate the development and progression of disease, and be able to identify promising therapeutic targets.

It would be tremendously valuable to be able to combine human genetic and genomic data gathered at scale, with disease-specific mechanistic and clinical studies to define the most compelling therapeutic targets, and to understand the cellular and physiological consequences of their perturbation. We would like to use human genetics (especially a series of alleles of diverse frequency, effects, and direction) to calibrate the relationship between the perturbation of a putative target and the resulting effects (both those that are desirable disease-modifying or mitigating effects and those with adverse outcomes). This would help to identify targets likely to have the best profile in terms of efficacy and safety. We also need to develop a range of disease-relevant models and functional assays and biomarkers that can provide essential support for both therapeutic development and clinical deployment.

The challenge lies in that these efforts are inherently non-scalable, at least at present. They typically examine processes, models and assays that are not generic, and that have to be carefully tailored to the phenotype and question of interest (sometimes referred to as a “final mile” problem). However, we believe there are opportunities to deliver more and more of the information that supports this goal through increasingly high-throughput, sophisticated, freely available datasets that can be shared across diseases to inform target identification and development. We also believe that there are opportunities for developing a more systematic perspective for target validation, in particular with respect to causal impact on disease onset or progression. This goal will address both key challenges in drug development—allowing both a higher proportion of successful targets and earlier failure for unsuccessful targets.

3. Foundational Resources

3.1 Human Cohorts.

A first major challenge is to gather as much data as possible about the effects of natural variation in the human population—to drive both scientific progress and clinical care. This challenge will require creating a wide range of foundational resources. In this initial version, we aim to pose key questions and lay out an initial work plan. Precise details remain to be filled in, as well as plans to achieve them.

(i) Create genomic characterization resources needed for any population. Deep characterization of genetic variation in a population is critical for any meaningful genetic studies.

We should **carefully define the extent of genetic characterization that should be obtained** for any population to be studied or served by clinical genetics, including (i) a collection of whole-genome sequences, with sufficient samples and high-enough quality (to detect SNPs and indels at a specified accuracy, detect heterozygous bases at a specified coverage, and allow imputation to a specified degree), (ii) a publically-available variant server (such as ExAC or gnomAD) allowing physicians and researchers to interpret variants found in a patient; (iii) well-designed genotyping arrays and publicly-available imputation server (until such time as genotyping is overtaken by whole-genome sequencing). We should then **ensure that these resources are created for any major population** to be studied and served by genetics.

(ii) Genotype existing collections. We should identify important existing disease-focused and population-based cohorts that have not yet been genetically analyzed and develop plans to ensure efficient sequencing and rapid data access.

(iii) Expand existing cohorts. Much larger human cohorts will be needed to understand human diseases—ideally including both disease-focused cohorts and biobanks linked to full medical records.

After reviewing existing disease-focused cohorts, we should **identify those diseases that would benefit from major expansion and the most efficient ways to increase a defined target number of patients**. For a subset of diseases, the number of cases should be large enough to provide good power for RVAS across most genes. The feasibility of using existing infrastructure for sample collection (e.g., in the US, collection systems for the *All of Us* project) should be explored.

After reviewing existing plans for biobanks, we should also **assess the prospects and barriers for creating biobanks**, and **identify the support that would help accelerate progress,** including laboratory and computational infrastructure**.**

To the greatest extent possible, cohorts should incorporate ongoing collection of clinical information by linking to medical records, to enable rich clinical characterization and routine longitudinal data. (They should also incorporate other kinds of information—including other records, such as demographic data that might shed light on environmental exposures; self-reported data from questionnaires; and data from new digital technologies, such as wearable devices.)

Where possible, cohorts should facilitate recall-by-genotype and recall-by-phenotype to enable biological follow-up studies.

(iv) Broaden studies to include all major populations. With the vast majority of genetic studies being in European-ancestry populations, there is an urgent need to include a much larger range of major populations and their subpopulations. (Machine learning methods can be useful in identifying cryptic sub-populations.)

After reviewing the issues, we should **develop an effective plan that would allow national and philanthropic funders to greatly expand the range of population studies** to expand the range of genetic discovery and to serve patients in various countries.

(v) Learn from special populations. Special populations, including founder populations and populations with high rates of consanguinity, provide the opportunity to gain genetic information that could not readily be learned in other ways—including discovering higher-frequency deleterious alleles and recessive loss-of-function phenotypes.

We should **identify the most promising special populations** (considering both genetic and feasibility issues) and **the support that would help accelerate progress.**

(vi) Create widely-accessible data platforms. Genetic analysis would be accelerated by the availability of high-quality, widely accessible data platforms that enable cloud-based data storage of human genetic data and cloud-based analysis with best-practices pipelines and analytical tools. Such platforms should eliminate unnecessary duplication of effort, including enabling immediate integration and harmonization of variant calls from large population cohorts and disease-focused cohorts, and should facilitate data sharing. Importantly, the data platforms should adhere to GA4GH standards.

We should identify the needs and any barriers to the creation and dissemination of such data platforms, including access by researchers in resource-poor settings.

(vii) Ensure sharing of summary statistics. It is crucial that the summary statistics for all GWAS be readily available to researchers. We should develop plans for a central repository of association data, with APIs for easy access, and policies to encourage and ensure deposit.

(viii) Enable federated analysis of individual-level data. We should develop approaches that allow federated analysis of individual-level data, while preserving security and privacy. Possible solutions include computer science techniques (such as secure multiparty computation, homomorphic encryption and other methods) and effective policies based on limited access among trusted parties. Attention will be needed toward relevant national and international law (such as Europe's General Data Protection Regulation). We also need ways to facilitate scientific collaboration among cohorts, including national biobanks.

(ix) Facilitate harmonization of phenotypes. Genetics can be used to assess consistency among different phenotypic definitions. For example, consistency can be assessed by (i) comparing the effect size estimates for genome-wide significant loci observed in a new cohort or (ii) assessing the genetic correlation between a new cohort and existing meta-analysis for the trait in question. Any significant differences can be probed by examining the pattern of genetic correlations with other traits.

(x) Enable sequencing of difficult regions and variants. Some regions of the human genome and some genetic variants remain difficult to sequence accurately. (Examples include human leukocyte antigen (HLA) genes at 6p21, killer-cell immunoglobulin-like receptor (KIR) genes at 19q13, Y chromosome variants and their haplogroups, and mitochondria variants and their haplogroups. These loci can have important effects on disease and can vary substantially across ancestry groups.) While these gaps are unlikely to represent rate-limiting steps in the understanding of common disease, we would ideally like to obtain complete and accurate coverage of the entire genome. New technologies are being developed, but they remain too expensive to deploy at large scale. We should encourage technologies that achieve long reads at high accuracy and low cost.

3.2 Analysis of Association Data.

Another major challenge is to be able to analyze human genetic data to extract as many insights as possible. Genetic mapping studies can do much more than identify individual disease-associated loci. Many creative approaches are being devised, and these developments will continue to be fueled by the growing availability of new kinds of data. In this initial version, we outline some key challenges for the analytical community.

(i) Sharpen GWAS signals. How can we best identify the variants responsible for GWAS signals—based on fine-mapping from larger datasets and haplotype structure, data from diverse populations, biological correlates, and other information?

(ii) Identify disease-relevant cell types. How can we best use the genome-wide effect-size vector from GWAS to identify the disease-relevant cell types? What sensitivity and resolution can we achieve with respect to cell types, states and contexts?

(iii) Identify disease-relevant cellular programs. Given a catalog of cellular programs (a set of coregulated genes), how can we best use the genome-wide effect-size vector from GWAS to identify the disease-relevant cell types?

(iv) Use disease-related phenotypes. How can we best exploit novel disease-related phenotypes to map and study diseases? Examples include family history (to increase the effective number of cases) or disease progression.

(v) Use pleiotropy across diverse phenotypes. How can we best use information GWAS data across many phenotypes to: annotate individual loci, variants, and haplotypes? partition a GWAS signal into latent factors? sharpen a GWAS signal by deriving latent traits with higher heritability? improve identification of relevant cell types and cellular processes?

(vi) Generate polygenic scores. How can we best generate polygenic scores? How do polygenic scores improve with increasing sample size? Can we substantially improve current methods? How can we combine polygenic scores with clinical information to increase accuracy and minimize biases? How can we avoid biases due to population and geographic stratification?

(vii) Use polygenic scores to identify disease-biology. To what extent and how best can we use polygenic scores to identify causal processes in presymptomatic individuals? Can we use polygenic scores for biomarkers as instruments to make causal inferences about outcome phenotypes, such as longevity, survival, clinical prognosis, and side effects (in effect, expanding Mendelian randomization from individual SNPs to polygenic scores)?

(viii) Use polygenic scores in clinical care. How should polygenic scores best be used in clinical care? How should polygenic scores be combined with other personal information (e.g., scores for cardiovascular disease with lipid levels) to yield accurate predictions?

(ix) Use polygenic scores in drug development. How can we best use polygenic scores to select patients for clinical trials?

(x) Understand selective forces. Can we best use the genome-wide distribution of effect sizes and allele frequencies (for both CVAS and RVAS) for various diseases and traits to make inferences about the selective forces and genetic architecture?

(xi) Understand contributions from the environment. How best to measure and characterize environmental exposures for mobile populations? How to consider life-stage specific genetic drivers of disease arising from an exposure at a critical window of susceptibility?

3.3 From Genetic Variants to Cellular Function.

Another major challenge is to be able to connect variants to function (V2F). The challenge entails: identifying the causal variants in each disease-associated locus and the disease-relevant cell type(s) in which they act, and identifying the immediate molecular consequences of the variants in a given cell type, such as the effect on an enhancer and on its target genes.

At present, these questions tend to be approached in an *ad hoc* manner. A more systematic approach will require creating comprehensive data resources, using both **observational** and **perturbational** approaches. Observational data can likely be applied to all human cell types, whereas perturbational approaches can likely be applied only to a limited set of cell types. Because it may not be possible to generate a complete look-up table in the foreseeable future, it will also be necessary to learn effective principles and rules that allow us to make high-quality inferences. In this initial version, we sketch some of the possible directions.

(i) Annotate functional elements in every cell type. A first step in analyzing a disease-associated locus is to identify how and where the causal variants might work by reference to functional elements in the region in various cell types (such as promoters, transcribed sequences, splice sites, enhancers, CTCF sites, etc.). Various projects (ENCODE, Epigenomic Roadmap, GTEx) have provided information at the level of cell lines and bulk tissue.

We now need to create **comprehensive maps of functional elements** for every human cell type. The single-cell methods and studies of the Human Cell Atlas are making it possible to comprehensively observe gene expression (by RNA-seq) and the open-chromatin regions (by ATAC-seq). For other epigenomic features, single-cell measurements are not yet feasible or not feasible at scale—but it may be possible to make high-quality inferences.

(ii) Map *cis*-regulatory QTLs in every cell type. For every common variant in the human genome, we would like to know whether it is associated with a *cis*-regulatory effect on gene expression (cis-eQTLs), chromatin structure (cis-cQTLs), or splicing (cis-sQTLS) in any human cell type. In principle, this can be done by applying single-cell analysis to cell types from a collection of individuals. (Effects can be detected by comparing expression across individuals of different genotypes (AA, AB, BB) or ideally allele-specific expression within individual heterozygotes (AB).)

We now need to create **genome-wide maps of *cis*-regulatory QTLs**. Initial efforts might concentrate on particular organs, to refine methods and define the appropriate scale.

(iii) Interpret enhancer function. We need to understand how an active enhancer functions in any given cell type—including which transcription factors (TFs) bind the enhancer and how genetic variants affect the binding in the cell type. This will likely require large-scale **perturbational experiments** and computational analysis in certain cell types, from which general rules can be derived to make about other cell types. Various approaches are available, including (i) genome-wide sequence analysis of enhancers to infer the roles of TFs, based on data within and across related cell types; (ii) genome editing in native enhancers; (iii) massively parallel reporter assays (MPRA) in heterologous settings, and (iv) machine learning. The utility of these approaches needs to be assessed in rigorous large-scale studies across many cell types.

(iv) Connect enhancers to promoters. We need to understand the rules that define the regulatory connections between enhancers and promoters in a given cell type—that is, which enhancers regulate which promoters and to what extent. As above, this will likely require large-scale **perturbational experiments** and computational analysis in certain cell types, from which general rules can be derived. Various approaches are available—including CRISPR inhibition (CRISPRi) of native enhancers and directed insertion of heterologous enhancers—but will need to be assessed at large scale.

(v) Assess protein-coding variants. We need generic assays to reliably determine whether and how a protein-coding variant in a gene alters cellular function; the assay should be applicable to all genes and at sufficient scale that it can be applied to all common variants and to any rare variant of interest. One powerful approach is to compare the effect of alleles on genome-wide expression across a sufficient number of cell types.

(vi) Catalog cellular programs. We need generic ways to infer the consequences of altering the expression level of any gene in any cell type. Generic assays are already available for altering a gene's expression (e.g., CRISPRi) and reading out its consequence on cellular gene expression (e.g., by RNA-seq), in those settings where human cells may be experimentally manipulated (such as cell lines and organoids).

We need much better ways to interpret the gene expression consequences in terms of meaningful cellular programs. We would ideally like to have a **comprehensive catalog of cellular programs**, consisting of the modules of gene expression triggered during development, physiology and disease. Such a catalog could aid us in interpreting the genetics of common disease—for example, by identifying cellular programs that correlate to the genome-wide effect-size vector for a disease and by clustering disease-associated genes for which expression changes trigger the same cellular programs.

Using machine learning, it may be possible to infer catalogs of cellular programs from the expression data that will be collected on billions of single cells in the coming years.

(vii) Identify 'missing' functional elements. While disease-associated variants are clearly enriched in known functional elements, the extent to which we are missing important classes of functional elements remains unclear. Careful analysis of the genome-wide distribution of disease-associated variants may shed light on this question, although it will be important to account for incomplete data about cell types.

3.4 Mechanisms and Medicine.

A fourth challenge is to establish a comprehensive understanding of the mechanisms involved in the onset and progression of disease, and to use these fundamental insights to drive the development of novel strategies for prevention and treatment. Existing approaches have typically relied on detailed characterization—involving disease-specific cellular and animal models based on bespoke assays—that have not easily lent themselves to high-throughput approaches.

However, as described above, translational efforts are now increasingly able to benefit from the foundational information provided by large-scale data generation in human genetics and cellular genomics. There are a number of additional activities where collaborative research can accelerate the translation of these high-throughput data for clinical benefit.

(i) Develop metrics of success to optimize target selection. There are currently no agreed-upon standards for defining progress toward connecting genetic signals with disease causation. We need frameworks that reflect the strength of such connections—for example, the extent to which mechanistic studies have "saturated" a process implicated in causation of a disease. Such metrics will also aid in evaluating and optimizing approaches for target selection.

(ii) Create a genetic dose-response portal. We need accessible tools that, through the aggregation of genetic, clinical, and functional data, provide a holistic description of the consequences of natural variation and experimental perturbation for each gene. This portal will particularly benefit from clinical studies performed in individuals with rare, large-effect genotypes. This information will support target discovery, calibrate the magnitude of anticipated therapeutic manipulation, help to anticipate adverse effects and toxicity, and identify putative biomarkers.

(iii) Develop common frameworks of disease architecture. The highly polygenic nature of common disease has implications for the ways in which we should interpret the mechanistic links between genetic signals and disease. We believe there is a need to develop common frameworks for mapping disease biology onto cellular and physiological processes.

(iv) Develop robust disease-specific models and functional assays. We need more sophisticated and robust cellular and tissue models of disease (including cell-lines, organoids, co-culture systems, and "organ-on-a-chip") together with approaches to scale these up for high-throughput interrogation. We also need disease-relevant functional assays to support multiple stages in target discovery and drug development, including high-throughput tools for characterizing variant and gene function, for the evaluation of therapeutic candidate molecules, and for conducting unbiased phenotypic screens. The collection and integration of extensive experimental data will, over time, enable the development of improved *in silico* models to support mechanistic inference regarding aspects of variant and gene function.

(v) Facilitate discovery of biomarkers. Large-scale, population-level, analyses of the proteome, metabolome and small non-coding RNA (in blood and other clinically relevant biofluids), across diverse populations and environmental exposures, are needed to support the direct and indirect (through Mendelian randomization approaches) discovery and characterization of clinical biomarkers. Actionable biomarkers that provide readouts related to target engagement, and pharmacodynamic predictors of efficacy and toxicity, as well as those that can capture risk and disease subtype, are an essential component of successful drug development.

(vi) Develop population-based approaches for personalized medicine research. There is a critical need to support large-scale recruitment of presymptomatic, at-risk, individuals with extensive baseline characterization (including polygenic scores), linkage to medical records, and consent for targeted follow-up (including genotype-based recall) able to support early biomarker detection, analysis of progression phenotypes, and interventional trials. There are also opportunities to use genetic and genomic data to assess the causal contributions of modifiable risk factors for disease (including aspects of lifestyle, the external environment and the microbiome) and thereby highlight strategies for disease prevention that can be used in both population-wide and targeted approaches.

(vii) Catalyze collaborative efforts across diseases. There is growing recognition, largely as a result of human genetics, that apparently disparate diseases often involve derangement of the same biological and cellular processes (examples include autophagy and fibrosis). There is a great deal to be gained by collaborative research to address aspects of biology that are relevant to multiple diseases, particularly those with substantial unmet clinical need. Repurposing approved or soon-to-be-approved drugs across traits based on genetic and biologic insights will accelerate patient benefits.

Chapter 3. Conclusions

In response to these goals, ICDA developed and released Recommendations (v1.0) based on the input of a diverse and engaged community of researchers, clinicians, funders, and policy experts.

The next step for ICDA will be to work with the scientific community to help implement these recommendations. We also welcome your input on how we might continue to refine and improve these recommendations at https://www.icda.bio/.

Contributors

ICDA Organizing Committee

Co-Chairs

Eric Lander	Broad Institute of MIT and Harvard
Cecilia Lindgren	University of Oxford

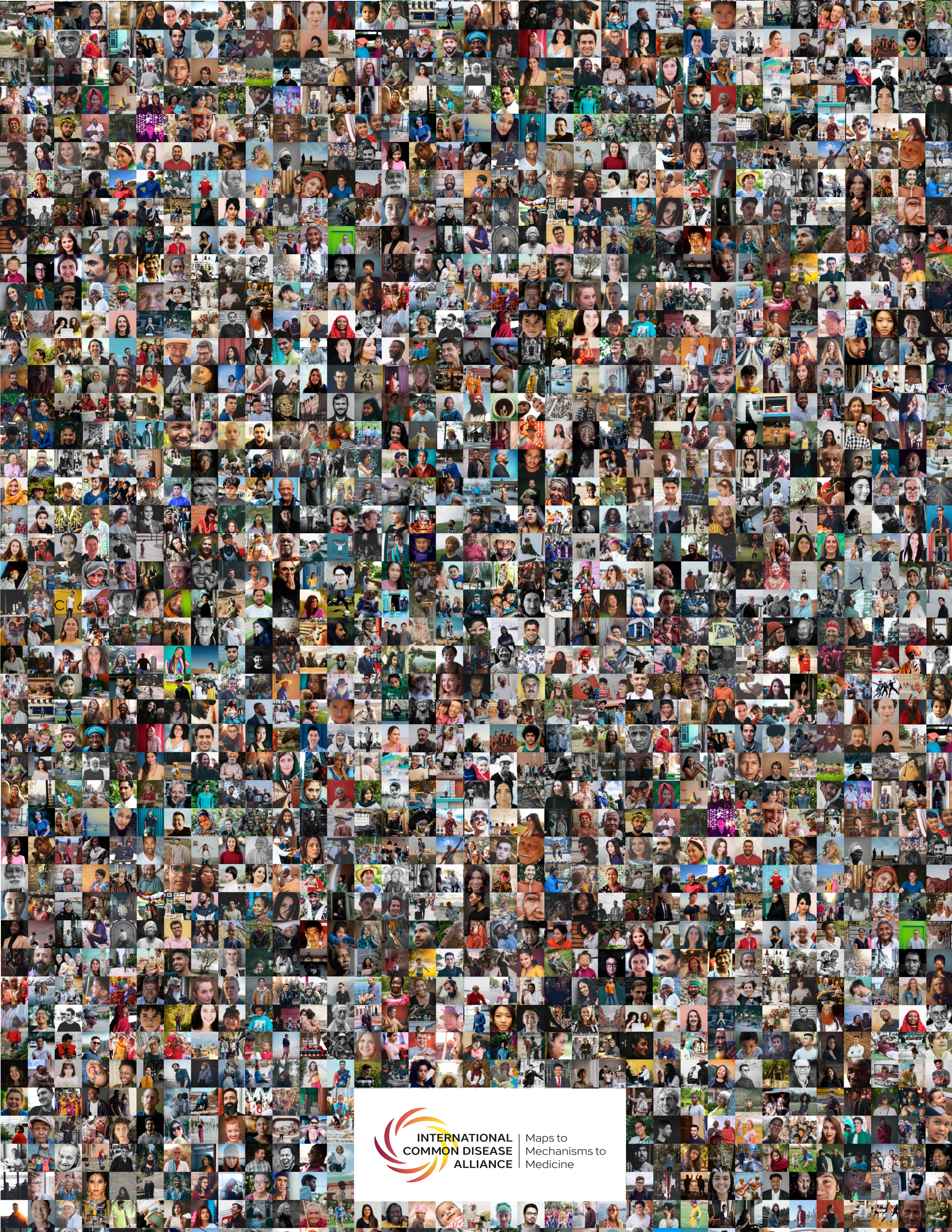
Executive Director

Rachel Liao	Broad Institute of MIT and Harvard

Members

Søren Brunak	University of Copenhagen
Judy Cho	Icahn School of Medicine at Mount Sinai
Rory Collins	University of Oxford
Nancy Cox	Vanderbilt University
Mark Daly	Institute of Molecular Medicine Finland (FIMM), University of Helsinki
George Davey Smith	University of Bristol
Emmanouil Dermitzakis	University of Geneva
Michael Dunn	Wellcome Trust
Lude Franke	University Medical Centre Groningen
David Glazer	Verily Life Sciences
Matthew Hurles	Wellcome Sanger Institute
Carolyn Hutter	National Human Genome Research Institute
Nancy Ip	Hong Kong University of Science and Technology
Sally John	Biogen
Tuuli Lappalainen	New York Genome Center, Columbia University
Partha Majumder	National Institute of Biomedical Genomics India
Mark McCarthy	Genentech
Andrés Moreno Estrada	National Laboratory of Genomics for Biodiversity Mexico
Benjamin Neale	Broad Institute of MIT and Harvard, Massachusetts General Hospital
Yukinori Okada	Osaka University Graduate School of Medicine
Helen Parkinson	EMBL-European Bioinformatics Institute
Charles Rotimi	National Human Genome Research Institute
Jay Shendure	University of Washington / Howard Hughes Medical Institute
Nicole Soranzo	Wellcome Sanger Institute

Kári Stefánsson	deCODE genetics
Patrick Sullivan	University of Northern Carolina, Karolinska Institutet
E Shyong Tai	National University of Singapore
Nicki Tiffin	University of Cape Town
Ricardo Verdugo	University of Chile
Cristen Willer	University of Michigan
Ambroise Wonkam	University of Cape Town
Unnur Þorsteinsdóttir	deCODE genetics



INTERNATIONAL
COMMON DISEASE
ALLIANCE

Maps to
Mechanisms to
Medicine