From Maps to Mechanisms to Medicine:

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

International Common Disease Alliance
White Paper

DRAFT v0.1 (for initial discussion)

September 23, 2019

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.

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. This early version (v0.1) is a partial draft intended 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. (The text currently lacks references and has some gaps to be filled.)

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

Chapter 3 (*not yet written*) will aim to (i) identify key projects, platforms and resources needed now to propel progress, and (ii) define clear and implementable plans to accomplish them.

Over the coming months, ICDA Working Groups will solicit input broadly from the ICDA community and propose specific recommendations. A revised version of this ICDA White Paper (v1.0) is planned for January 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 make-up 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 affect roughly 1 in every *** 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 a 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 it 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, with ~67,000 cases and ~94,000 controls, reveals 245 loci (PGC Schizophrenia Working Group). 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
Add other diseases/traits	

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 common diseases. 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 < 2%, 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 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 disease, 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 characterisation, 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 1 million) — and by some major medical centers (with early leaders such as Vanderbilt's BioVu, Geisinger 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. 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.

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 (i.e., 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 (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 preferably hundreds of thousands of cases (Supp. Table 2). 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 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 is 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 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 elements. (Likelihood analysis suggested ~3% of bases in ~2% of these elements might be pathogenic.)

For RVAS in non-coding regions to become practical, huge leaps would be required in the precision in identifying 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 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.

When one has a large cohort with a wide range of phenotypes, one 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, one 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.

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. These 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, to enrich for events and thereby decrease the required sample size.

For an individual's polygenic risk scores to have maximal accuracy and precision, it 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 be addressed rapidly.

In addition to improving methods to derive scores from CVAS and expand the representation of ancestries, integrating CVAS risk with RVAS risk will be essential to delivering accurate and complete genetic risk scores. Although rare variants contributes less on a population level, they often contribute 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.

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. The effect sizes observed in GWAS (that is, the genome-wide effect size vector) 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

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 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 80% 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, GWAS 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.

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—e.g., 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 (ref: Miga et al. https://doi.org/10.1101/735928).

3.2 Catalogues 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 dissemination 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, using such methods as RVIS, MPC, and louef. 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 been 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 (ref). 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 has 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 epigenomics 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.

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 the 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 of 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 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 more restricted than effects observed when 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 and clinical data to best stratify risk and disease subtype in ways that that may have clinical utility (“precision medicine”). For example, the capacity to stratify patients can be critical for Phase 2 studies (e.g. picking out 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 maximising efficacy, minimizing adverse events, and/or addressing unmet clinical need).

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 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-Meyers 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 100,000 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 and then to disease treatment. Even the first step has been accomplished only in a small fraction of cases to date.

To make 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-based data platforms, enabling scientific communities to store and analyse massive datasets in commonsettings;
- 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 rapidly move 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 variants 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 the efficiency of clinical trials 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 causal variant(s). We want ultimately 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.

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 program.) 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, 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.

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-based data storage of human genetic data and cloud-based 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 clinical services are delivered at 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 most 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 year, 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. The September 2019 meeting near Washington, DC will be the official launch of the ICDA.

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

- **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;
- **define current barriers to progress,** including scientific, technological, computational and organizational obstacles;
- **identify needs and opportunities for new projects** 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 to drive progress**, including key knowledge, datasets, experimental technologies, computational platforms, and frameworks for data sharing and data harmonization;
- **develop white papers proposing effective plans** to enable these solutions;
- **coordinate with funders** to ensure 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 undertake 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; developing ethical frameworks for truly global collaboration; and others?

Role of partnerships. Partnerships of many kinds will be critical for success, involving clinicians and researchers, epidemiologists and geneticists, 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 will aim 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 datasets, experimental technologies, analytical methods, computational tools, and data platform. 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 implementable plans** for how these things might best be accomplished.

This initial draft (v0.1) makes a start at the first three items above. It was developed with the aim of beginning discussion within the ICDA community about specific plans that would best address the community's needs.

As a next step, the ICDA Working Groups will invite input about proposed goals, directions and plans from the entire community. The first full draft of this White Paper (v1.0) is planned for January 2020.

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 consequence 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 (~1.5×10^{10}) and mutation rate per nucleotide per generation (~1.3×10^{-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 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 for 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 precise which transcription factors are binding at each enhancers 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 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 the 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, series of alleles of diverse frequency, effects and direction) to calibrate the relationship between the perturbation of 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 these efforts are inherently non-scalable, at least at present. They typically involve 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 publically-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 analysed 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 cases 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 linkage 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 to 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 range from the HLA genes on chromosome 6 to centromeres and complex pericentromeric repeats.) 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 narrow the variants responsible for GWAS signals—based on fine-mapping based on 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-wise 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-wise 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 presymptomatic patients 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 best can we use polygenic scores best be used 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?

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 variant(s) in each disease-associated locus and the disease-relevant cell type(s) in which it acts, 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 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 allow make inferences 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 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, for each gene, a holistic description of the consequences of natural variation and experimental perturbation. 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 for 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 and metabolome (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 involved 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. Next Steps: Developing Plans and Recommendations

The purpose of this initial version of the 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.

The ICDA community must now define:

(i) the key projects, platforms and resources needed now to propel progress, and
(ii) clearly defined and implementable plans to accomplish them.

Over the next several months, ICDA Working Groups will solicit input broadly and make specific recommendations. A revised version of this ICDA White Paper (v1.0), including a first set of recommendations, is planned for January 2020.

ICDA's goal is that these recommendations—intended for researchers and funders—will be useful in helping the scientific community converge on common goals for the next phase of human genetics.

More broadly, ICDA aims to serve as a scientific forum for ongoing discussions, including scientific meetings, about the exciting work ahead.