AMP T2D RESEARCH PLAN

Accelerating Medicines Partnership Type 2 Diabetes Program Executive Summary

The Accelerating Medicines Partnership (AMP), formerly the Target Validation Consortium, is a pre-competitive collaboration among government, academia and industry, to harness collective capabilities, scale and resources toward improving current efforts to develop new therapies for complex, heterogeneous diseases. The focus of the partnership is on doing the research necessary to understand these diseases more fully, identifying the right targets to pursue for drug therapy, and thereby accelerating the ability to bring new medicines to patients in these diseases.

The original joint research plans for three therapeutic areas – Type 2 Diabetes, Alzheimer's Disease, and Rheumatoid Arthritis, Lupus & Related Autoimmune Disorders – were drafted in the spring of 2013. In the fall of 2013, AMP completed revising the research plans through a series of meetings of the joint Steering Committees, including representatives from the following participating companies: AbbVie, Biogen Idec, Bristol-Meyers Squibb, GlaxoSmithKline, Johnson & Johnson, Lilly, Merck, Pfizer, Sanofi, and Takeda as well as members from government and academia.

The research plan outlined below represents a revised version of the original white paper, which has been approved by all members of the Type 2 Diabetes (T2D) Steering Committee, to ensure that the priorities and scientific agendas of each participating stakeholder are addressed effectively through the research plan. The next few pages will briefly describe the research objectives, timeline, milestones and budget as well as the overall governance for this disease area. Further detail on each section can be found in the attached detailed research plans.

I. Problem Statement and Value Proposition

Type 2 diabetes (T2D) currently affects about 26 million people in the US and over 170 million people worldwide with the prevalence increasing rapidly and the consequences including macrovascular disease (e.g., coronary artery disease, ischemic stroke, and peripheral vascular disease) as well as microvascular disease (e.g., diabetic retinopathy and nephropathy). While there are a number of approved T2D therapies on the market, there remains a major unmet medical need, as no therapy has been shown to achieve long-term reversal of the progression of hyperglycemia, or to prevent complications. The available therapies remain limited in part because the pathophysiology of insulin resistance, beta cell failure, and progression to T2D, remain poorly understood. Studies of humans with monogenic disorders of insulin resistance and T2D as well as unbiased genome-wide studies of humans with T2D have provided some insight into pathways that are causal. However, the majority of loci harbor genes that have no previously known biological relationship to the pathogenesis of T2D, indicating how little is understood about the root causes of this disease in human subjects.

Given the complex and intersecting pathways that control glucose homeostasis and energy balance, and the lack of clinical validity of existing cell and animal models, validation of drug targets for T2D has been challenging. One of the more promising approaches is to take advantage of human genetics to validate drug targets. Loss of function (LoF) or gain of function (GoF) gene variants that have large effects on T2D and related phenotypes can unveil relevant mechanisms and pathways and validate drug targets. There are proof-of-principle examples in the diabetes literature to support such an approach. Insights gleaned from these genetic "experiments of nature" are informed by hypothesis-driven phenotyping of a relatively small number of patients with mutations of large effect on T2D risk and related phenotypes.

AMP's T2D research plan aims to use human genetics as a powerful approach to validate targets *in vivo* in the human population, exploiting experiments of nature that perturb protein function. Of particular value are mutations of known molecular effect (e.g., loss of function) that result in a desirable clinical outcome (e.g., protection from disease) without adverse consequences. To discover, validate, and characterize a gene's potential for human target validation, data are typically needed from multiple study designs, access to samples from large or special collections, measurement of multiple phenotypes, and study of multiple variants in the same gene. That is why this project is focused on the systematic aggregation of existing genotype-phenotype data for T2D, related traits, and its complications, the generation of a large amount of new genotype data through extensive targeted re- sequencing of carefully selected samples, and as an aspirational goal, the detailed phenotypic characterization of carefully selected individuals bearing genotypes of high interest. Therefore, this project builds on the extensive amount of work that has already been done and fills a key niche that lies after target identification but before a commitment to pursue a specific target. This work will contribute to further deconstruction of the pathophysiology of T2D, and potentially of its major complications, including macrovascular and microvascular disease.

II. Project Overview and Specific Aims

The overall strategy of the diabetes project is to provide access to high-quality human genetic and phenotype data that will allow the evaluation of the efficacy and safety of potential therapeutic targets for diabetes, and its complications, and thereby inform the drug development pipeline. Program A describes the ways in which available and emerging human genetic data can be harnessed to this end. Program B focuses on the generation of new human genetic and genomic data for targets of particular interest.

Program A is the creation of a Knowledge Portal that researchers can use to identify relationships between sequence variation in potential targets in the genome and risk or protection from T2D, cardiovascular and kidney disease risk in T2D patients and related intermediate metabolic endpoints. An infrastructure to aggregate available genome sequence and phenotype data in T2D and cardiovascular and kidney disease in T2D, bringing together an array of data on samples characterized with both sequence and relevant phenotype data, will be established. Automated analytical methods and query tools will be deployed to provide the clearest and most interpretable answers about the relationships between gene function and diabetes related phenotypes. A key element of this program is the inclusion of data fields related to clinical sequelae as this is a major gap in our ability to progress our understanding in this highly relevant area. The database thus created will be used by the pharmaceutical industry and academic researchers and clinicians to allow the following types of hypotheses to be tested:

- **Phenotype-based queries:** Genetic variation, which protects from or contributes to T2D risk, is associated with variability in T2D-related traits or impacts the risk of T2D patients developing cardiovascular or kidney disease?
- **Gene- or pathway-based queries:** What genetic variation exists within a target or pathway of interest and is this variation associated with an increased or decreased risk of T2D, or impacts the risk of cardiovascular or kidney disease in T2D?
- **Variant-based queries:** What are the clinical, biochemical, expression quantitative trait loci, and epigenetic phenotypes, associated with a given variant? **Subset queries:** Are results consistent across ancestry groups and across studies?

Program A is expected to run over 5 years.

Program B focuses on the generation of new human genetic and genomic data for targets of particular interest (those for which the existing human genetic and genomic data available through the Knowledge Portal are insufficient to permit a robust "go/no go" decision) through "deep" genetics. As described below, an initial effort focused on the development of T2D itself, will be extended to provide equivalent insights into target validation for the complications of diabetes. For any given target, we expect Program B to generate data that contributes to answering five key questions:

- Is there evidence that perturbation of a target's function leads to a change in T2D status (diabetes-related quantitative metabolic traits or risk of diabetes complications) consistent with the expected outcome of therapeutic modulation?
- Is the desirable therapeutic modulation to be achieved through LoF or GoF?
- Is there evidence that perturbation of a target's function leads to "on target" adverse risk effects that would compromise its value as a therapeutic target?
- Does human genetics or genomics provide insight into the mechanism of action?
- Can human genetics identify individuals carrying high value alleles (e.g., rare variants of large effect) of interest for call back studies described in Program C?

The "genetic [or sequence] targets" of this endeavor will be defined by the consortium stakeholders and will comprise the set of genes encoding potential therapeutic targets for which more intensive human genetics validation is deemed to offer value. The focus would be on target validation with respect to T2D, but gene targets of particular interest would also be sequenced in case-control samples of coronary heart disease given the immense interest in knowing whether targeting a particular gene for T2D may reduce (or increase) the risk of coronary heart disease. Targeted analysis of genes of interest will be conducted across a wide range of samples using standardized reagents and protocols, with (largely, but perhaps not exclusively) centralized data generation and analysis. For most of the above purposes, a combined data set exceeding 100,000 to 150,000 individuals would be feasible and well-powered under a range of realistic genetic models**.**

Program B is expected to run over 5 years.

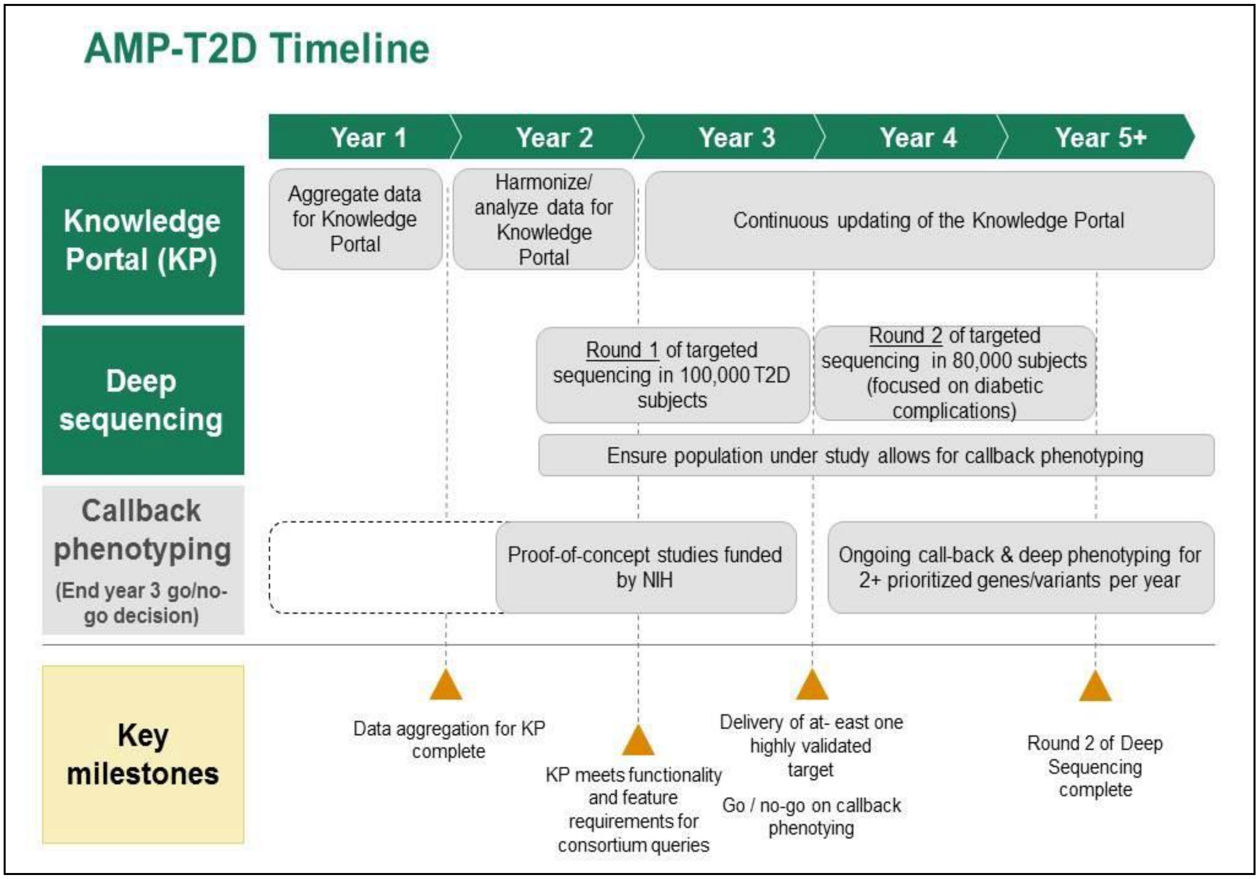
This overarching T2D effort is ideally suited for the collaborative AMP effort. The amount of genetic data generated over the past years in cohorts enriched with clinical information on diabetes, cardiovascular disease and their associated metabolic intermediate endpoints, is massive and unparalleled across other disease areas. In order to derive informative analyses from these studies, large sample sizes are required. Several groups have already formed alliances and are sharing data to allow for these integrated analyses, but there is currently no effort that has attempted to bring together the numbers proposed in this research plan. Moreover, in spite of existing funding requirements to deposit data into a centralized database upon completion of research, data are not readily available to investigators outside of the already formed alliances, which are operating now. That is why the need to pool available data into an accurate and easily accessible integrated database requires a consortium. Additionally, the plan described below requires large numbers of subjects for targeted sequencing, which makes it essential for multiple groups to work together to pool samples for sequencing and analysis. In summary, achieving the scope and scale required in this plan can only be effectively accomplished through a partnership, as described below.

III. Project Management and Project Funding Mechanisms

The Steering Committee for T2D will be comprised of representatives from participating companies as well as members from government, academia, and non-profit organizations, and will operate under the direction of the overall AMP Executive Committee (EC), comprised of 3-4 leaders each from industry and NIH, as well as a representative each from FDA, academia, and the patient advocacy sector. The EC is in turn advised by an Extended Executive Committee comprised of R&D heads of companies involved in the partnership. The T2D Steering Committee is responsible for defining the research agenda and project plan, for review of ongoing projects, and for the detailed assessment of milestones. The EC will also review the assessment of milestones and any revision to the project plan that results from a “No-go” assessment that some element of the current plan is not feasible.

IV. Timeline and Milestones

This integrated project, involving Programs A and B, is currently structured as a five-year initiative, in which Program A is anticipated to begin in the summer of 2014 with the first part of that year used to create a Request for Proposals at FNIH in parallel with the Request for Applications that will be used at NIH, to jointly select the institution(s) at which the knowledge portal will be housed. For program B, starting in the summer of 2014, an inventory of the needed samples and contracts, associated with their transfer will be put in place to allow the launch of Program B in 2015. From that point forward, the two programs will run in parallel until their end in mid-2019 as described on the figure below. The common AMP T2D Steering Committee for the two programs will ensure that substantial and effective scientific and logistical interactions occur between NIH- and FNIH-sponsored AMP projects within a fully integrated and collaborative AMP T2D consortium.



AMP-T2D Timeline

Year 1 | Year 2 | Year 3 | Year 4 | Year 5+

Knowledge Portal (KP)

Aggregate data for Knowledge Portal

Harmonize/ analyze data for Knowledge Portal

Continuous updating of the Knowledge Portal

Deep sequencing

Round 1 of targeted sequencing in 100,000 T2D subjects

Round 2 of targeted sequencing in 80,000 subjects (focused on diabetic complications)

Ensure population under study allows for callback phenotyping

Callback phenotyping
(End year 3 go/no-go decision)

Proof-of-concept studies funded by NIH

Ongoing call-back & deep phenotyping for 2+ prioritized genes/variants per year

Key milestones

Data aggregation for KP complete

KP meets functionality and feature requirements for consortium queries

Delivery of at- east one highly validated target

Go / no-go on callback phenotying

Round 2 of Deep Sequencing complete

Detailed milestones and go / no-go milestones are listed in the detailed research plan.

Contents

Section 0: Disease Context and Case for Action

Rationale

Type 2 Diabetes (T2D) currently affects 26 million people in the US and over 170 million people worldwide and the prevalence is increasing rapidly. The consequences of T2D include macrovascular disease (including coronary artery disease, ischemic stroke, and peripheral vascular disease) as well as microvascular disease (including retinopathy and nephropathy). Average lifespan in persons with T2D is shortened by 5-10 years. The costs of medical treatment of T2D and its complications in the US are $116 billion per year. While there are a number of approved therapies on the market for T2D, there remains a major unmet medical need as no therapy has been shown to achieve long-term reversal of the progression of hyperglycemia, nor to prevent complications. Many patients on even 2 or 3 drugs fail to achieve acceptable glycemic control, and many patients ultimately require insulin therapy. Furthermore, the macrovascular and microvascular complications of T2D represent major unmet medical needs and present opportunities for novel therapeutic development.

"State of the science"

In simple terms, T2D is characterized both by insulin resistance and the inability of the pancreatic beta cell to match the demand for insulin, beta cell failure, and insufficient insulin secretion to maintain normoglycemia. Insulin resistance is itself associated with a variety of other phenotypes, including dyslipidemia, hypertension, inflammation, hepatic steatosis, and atherosclerosis, that are co-morbidities of T2D as well. The pathophysiology of insulin resistance, beta cell failure and progression to T2D remains poorly understood. Studies of humans with monogenic disorders of insulin resistance and T2D as well as unbiased genome-wide studies of humans with T2D have provided some insight into pathways that are causal. For example, a handful of genes have been found to cause monogenic insulin resistance; the genes that cause monogenic forms of type 2 diabetes (MODY) act in the beta cell and the liver. Common variant genome-wide studies have yielded > 70 loci that are genome-wide significantly associated with T2D and many others with related glycemic traits. A substantial subset of these loci harbor genes that appear to be involved in the regulation of insulin secretion by the beta cell. However, the majority of loci harbor genes that have no previously known biological relationship to the pathogenesis of T2D, indicating how little we understand the root causes of T2D in human subjects.

Current research initiatives intersect with and amplify this proposal but are not directly overlapping. Over the past 5 years the human genetics community has collected human genetic and clinical data at a vast scale, and self-assembled into highly interactive collaborative groups. By the end of 2013, the field of T2D genetics will have collected data on a scale not available for any other disease: n>100,000 individuals genotyped genome-wide for common variants (GWAS), n>75,000 genotyped for low frequency coding variants using the exome chip, n=19,000 sequenced deeply in exons for rare and private variants, and n>4,000 whole genome sequenced at high coverage. Genetic analysis of myocardial infarction (of which are large subset have T2D) is similarly racing ahead: n>100,000 individuals genotyped for GWAS and for exome chip; and by 2014 n=14,000 exome sequenced for premature myocardial infarction.

We emphasize that the current proposal would leverage and take advantage of this investment, funding initiatives that would substantially move the field forward with regard to target validation for T2D and its complications.

Problem statement

Given the complex and intersecting pathways that control glucose homeostasis and energy balance, and the lack of clinical validity of existing cell and animal models, validation of T2D drug targets has been challenging. While some drug targets identified in non-human models have successfully translated to humans, the vast majority have failed due to the relative importance of the target in humans, pathway redundancies, and "on-target" adverse effects. One of the more promising approaches is to take advantage of human genetics to validate drug targets. Loss of function (LoF) or gain of function (GoF) gene variants that have large effects on T2D and related phenotypes can unveil relevant mechanisms and pathways and validate drug targets. There are proof-of-principle examples in the diabetes literature to support such an approach. These include rare LoF mutations in *ABCC8* and *KCNJ11* (the target for sulfonylureas) that cause neonatal diabetes, LoF mutations in *SGLT2* (the target for the new SGLT2 inhibitors) that causes a benign form of familial renal glucosuria, and even rare LoF mutations in *INSR* that causes type A syndrome of insulin resistance, and the insulin gene itself. Insights gleaned from these genetic "experiments of nature" were informed by hypothesis-driven phenotyping of a relatively small number of patients with mutations of large-effect on T2D risk and related phenotypes.

The proposed program in T2D aims to use human genetics as a powerful approach to validate targets *in vivo* in the human population, exploiting experiments of nature that perturb protein function. Of particular value are mutations of known molecular effect (*e.g.*, loss of function) that result in a desirable clinical outcome (*e.g.*, protection from disease) without adverse consequences. Examples of this approach include CCR5 (protection from HIV infection) and PCSK9 (reduction in LDL cholesterol and coronary disease risk). However, to discover, validate, and characterize a gene's potential for human target validation, data are typically needed from multiple study designs, access to samples from large or special collections, measurement of multiple phenotypes, and study of multiple variants in the same gene. Our proposal is focused on the systematic aggregation of existing genotype-phenotype data for T2D, related traits, and its complications, the generation of a large amount of new genotype data through extensive targeted sequencing of carefully selected samples, and as an aspirational goal, the detailed phenotypic characterization of carefully selected individuals bearing genotypes of high interest. This proposal builds on the extensive amount of work that has already been done and fills a key niche that lies after target identification but before a commitment to pursue a specific target, namely target validation space using human genetics as the primary tool. In the process, this work will contribute to further deconstruction of the pathophysiology of T2D, and potentially of its major complications, including macrovascular and microvascular disease.

Section I: Project Overview

Project goal and specific hypotheses to be tested

The majority of new chemical entities that enter human clinical trials fail, often due to lack of efficacy. Often, the problem lies not in the molecule or trial design, but in the choice of target.

That is, the molecule acts as intended, but the therapeutic hypothesis does not apply in patients. Leveraged against the tremendous cost of each late-stage failure, an effective up-front investment in improved target validation (and invalidation) should result in substantial returns downstream. By exploiting experiments of nature that perturb protein function we can provide insight into the clinical consequences of a given gene / pathway perturbation prior to development of a therapeutic agent. Particularly valuable are mutations of known molecular effect (*e.g.,* loss of function) that result in a desirable clinical outcome (*e.g.*, protection from disease) without adverse consequences. Examples of this approach include CXCR4/ CCR5 and protection from HIV infection and PCSK9 and reduction in LDL to prevent coronary disease.

This project is based on the hypothesis that naturally occurring genetic variation in humans linked to high quality phenotype data provides the best opportunity to validate new therapeutic targets for T2D and its complications. The overall goal of the project is to provide access to high-quality human genetic data that will allow the evaluation of the efficacy and safety of potential therapeutic targets for diabetes and thereby inform the drug development pipeline. We propose three interconnected programs to accomplish this goal.

Overview of the experimental plan

Our proposed approach is motivated by the experience that successful efforts to discover, validate, and characterize a gene's potential for human target validation nearly always draw on multiple study designs, access to samples from large or special collections, robust measurement of multiple phenotypic assessments, and require multiple variants in the same gene. The overall strategy of the diabetes project is to provide access to high-quality human genetic and phenotypic data that will allow the evaluation of the efficacy and safety of potential therapeutic targets for diabetes, and its complications, and thereby inform the drug development pipeline. Within Program A, we describe the ways in which we can harness available, and emerging, human genetic data to this end. In Program B, we focus on the generation of new human genetic and genomic data for targets of particular interest. In Program C, we explain how detailed physiological and genomic examination of individuals carrying alleles of particular interest can support this objective.

Program A is the creation of a knowledge portal that researchers can use to identify relationships between sequence variation in potential targets in the genome and risk or protection from type 2 diabetes (T2D), cardiovascular and kidney disease risk in T2D patients and related intermediate metabolic endpoints. We will establish an infrastructure to aggregate available genome sequence and phenotype data in T2D and cardiovascular and kidney disease in T2D, bring together an array of data on samples characterized with both sequence and relevant phenotype data, deploy automated analytical methods and query tools to provide the clearest and most interpretable answers about the relationships between gene function and diabetes related phenotypes. A key element of this program is the inclusion of data fields related to clinical sequelae as this is a major gap in our ability to progress our understanding in this highly relevant area. The database thus created will be used by industry and academic scientists to allow the following types of hypotheses to be tested:

- **Phenotype-based queries:** How is genetic variation which protects from or contributes to T2D risk associated with variability in T2D-related traits or the risk of T2D patients developing cardiovascular or kidney disease?

- **Gene- or pathway-based queries:** What genetic variation exists within a target or pathway of interest and is this variation associated with an increased or decreased risk of T2D, or impacts the risk of cardiovascular or kidney disease in T2D?
- **Variant-based queries:** What are the clinical, biochemical, expression quantitative trait loci, and epigenetic phenotypes associated with a given variant?
- **Subset queries:** Are results consistent across ancestry groups and across studies?

Program A is expected to run over 5 years.

Program B will focus on the generation of new human genetic and genomic data for targets of particular interest (those for which the existing human genetic and genomic data available through the Knowledge Portal are insufficient to permit a robust "go/no go" decision) through what we have termed "deep" genetics. As we describe below, an initial effort focused on the development of T2D itself, will be extended to provide equivalent insights into target validation for the complications of diabetes. For any given target, we expect Program B to generate data that contributes to answering five key questions:

- Is there evidence that perturbation of a target's function leads to a change in T2D status (diabetes-related quantitative metabolic traits or risk of diabetes complications) consistent with the expected outcome of therapeutic modulation?
- Is the desirable therapeutic modulation to be achieved through LoF or GoF?
- Is there evidence that perturbation of a target's function leads to "on target" adverse risk effects that would compromise its value as a therapeutic target?
- Does human genetics or genomics provide insight into the mechanism of action?
- Can human genetics identify individuals carrying high value alleles (*e.g.,* rare variants of large effect) of interest for call back studies described in Program C?

The "genetic [or sequence] targets" of this endeavor will be defined primarily by those biopharma companies participating in the TVC and will comprise the set of genes encoding potential therapeutic targets for which more intensive human genetics validation is deemed to offer value. The focus would be on target validation with respect to T2D, but gene targets of particular interest would also be sequenced in case-control samples of coronary heart disease given the immense interest in knowing whether targeting a particular gene for T2D may reduce (or increase) the risk of coronary heart disease.

Targeted analysis of genes of interest will be conducted across a wide range of samples using standardized reagents and protocols, with (largely, but perhaps not exclusively) centralized data generation and analysis. For most of the above purposes, a combined data set exceeding 100,000 to 150,000 individuals would be feasible and well-powered under a range of realistic genetic models**.** Program B is expected to run over 5 years.

Program C is an aspirational goal of the partnership. We will focus on deep phenotyping of carefully selected individuals with rare loss of function (LoF) or gain of function (GoF) gene variants that have large effects on T2D and related phenotypes. There are a number of proof-of-principle examples in the diabetes literature to support such an approach. The overall goal of this program is to extend this human genomics approach to target validation by systematically identifying patients with novel LoF and GoF variants that cause diabetes for “callback” in order to perform deep hypothesis-driven phenotyping to understand mechanisms and pathways and to validate novel T2D targets. Deep phenotyping will also provide the opportunity to evaluate patients for “on-target” adverse events that might preclude the target from further consideration. A similar approach may be used to validate targets for prevention/delay/treatment of diabetic complications such as nephropathy, retinopathy, neuropathy, and cardiovascular disease. A particular challenge for this initiative will be the actual experimental design, both in terms of subject burden and selection of appropriate control subjects. Experimental validation of a particular target may require a comprehensive approach, including animal studies and studies in human cell lines. This program may require a long term commitment of funds prior to an assessment of its utility and a go, no go decision. Pilot experiments using this approach are now underway in several academic labs and will need to be evaluated for impact later in the course of the partnership effort. The potential impact of this program will be re-evaluated when sufficient data is available.

Benefits of the AMP T2D Program

The greatest value is to aid in the identification and validation of therapeutic targets for diabetes. There are very few novel targets in diabetes. Many companies with efforts in this disease area work on the same targets and often end up failing either because the target does not have efficacy in humans or there are safety issues. The fact that so many companies work on the same ineffective targets reflects our lack of approaches to identifying targets with human disease relevance. The failures based on lack of efficacy are in large part due to the poor predictive value of preclinical animal models. In addition, there is a strong interest in industry to treat not just glucose and hemoglobin A1C, but the end-organ manifestations of diabetes such as cardiovascular disease. The preclinical models for end-organ disease are even less relevant than the more proximal endpoints. The need to have translational tools such as genetics in humans with diabetes and cardiovascular disease would greatly help industry focus on more relevant targets.

Admittedly, the database is likely to serve as a starting point or as a piece to an overall validation package for a new diabetes target, but the opportunity to start with a target that instantly has human relevance is a large step forward for industry. Once an association can be identified, there are a number of potential next steps that can further validate the target and help with biomarkers and patient segmentation. These include additional genetic studies to extend the observation (as described in Project B) and call-back studies for deep phenotyping of individuals with well-characterized variation in the gene of interest (as described in Project C). Scientists in industry with rare exception do not have access to genetic data, and those who do have limited access through collaborations with individual partners (*e.g.*, Amgen-DeCode) or through single cohorts (*e.g.*, GSK), so access to this database could have significant impact on our ability to successfully develop the next generation of diabetes therapies. The ability to access human data through this portal will greatly improve the translational relevance of the science performed in academic centers, which will have a direct effect on industry. Finally, the value to patients is obvious. Diabetes is a growing epidemic, and despite outstanding preventive therapies, cardiovascular disease represents the number one killer worldwide. A better understanding of the drivers of T2D and T2D-associated cardiovascular disease in humans will ultimately lead to a next generation of effective therapies that will be targeted to those patient segments who are most likely to benefit.

Need for a partnership

These projects are ideally suited for and require a partnership. The amount of genetic data generated over the past several years in cohorts enriched with clinical information on diabetes, cardiovascular disease and their associated metabolic intermediate endpoints is massive and unparalleled across other disease areas. It is likely that more genetic data has been amassed in metabolic diseases and related cardiovascular diseases then all other diseases combined. In order to derive informative analyses from these studies, large sample sizes are required. Several groups have formed alliances and are sharing data to allow for these deep analyses, but there is no single effort that has attempted to bring together the numbers proposed in Program A. As importantly, in spite of existing funding requirements to deposit data into a centralized database upon completion of research, data are not readily available to investigators outside of these alliances, in industry and often times to members within the alliance. The need to pool these data into an accurate and easily accessible database requires a partnership given the necessary collaboration and costs. For Program B, the numbers of subjects required for targeted sequencing means that it is essential for multiple groups to work together to pool samples for sequencing and analysis. For Program C, finding the most informative individuals who, for example, have rare loss-of-function mutations in both alleles of a gene of interest, will require collaboration among a large number of groups. Furthermore, it will be necessary to harmonize phenotyping efforts in order to generate the most robust phenotype data in these rare, highly valuable patients. In summary, this proposal of these three ambitious interconnected Programs can only be accomplished through a coordinated partnership effort.

Section II: Scientific Design

The overall strategy of the diabetes project is to integrate, analyze and provide access to high-quality human genetic and linked phenotype data that will allow the evaluation of the efficacy and safety of potential therapeutic targets for diabetes and its complications, and thereby inform the drug development pipeline. Within **Program A**, we describe the ways in which we can maximize the value and accessibility of available, and emerging, human genetic data to this end. In **Program B**, we focus on the deep characterization of targets of particular interest to drug discovery. In **Program C**, we explain as an aspirational goal how detailed physiological and genomic examination of individuals carrying alleles of particular interest can inform knowledge of the physiological effects of pathway perturbations in humans.

Program A: A Knowledge Portal to enable target validation based on human genetics of T2D and its complications

Background

Program A aims to generalize a powerful approach to validate targets *in vivo* in the human population: combining, interrogating and making accessible information on human sequence variation in combination with rich phenotype data. This approach has been enabled by the collection of human genetic and clinical data at a large scale, and by the self-assembly of investigators into highly interactive collaborative groups. By the end of 2013, the field of T2D genetics will have collected n>100,000 individuals genotyped genome-wide for common variants (GWAS), n>75,000 genotyped for low frequency coding variants using the exome chip, n>19,000 sequenced deeply in exons for rare and private variants, and n>4,000 whole genome sequenced at high coverage. Genetic analysis of myocardial infarction is similarly racing ahead: n>100,000 individuals genotyped for GWAS and for exome chip; and by 2014 n>14,000 exome sequenced for premature myocardial infarction.

Program A aims to create a Knowledge Portal that will reveal relationships between sequence variation in potential targets (genome-wide) and risk of T2D, related quantitative measures, and cardiovascular disease and major microvascular complications in T2D patients. We will establish an infrastructure to aggregate available data on genome sequence and phenotype in T2D subjects with and without cardiovascular disease, bring together an array of data on samples characterized with both sequence and relevant phenotype data, deploy automated analytical methods and query tools to provide the clearest and most interpretable answers about the relationships between gene function and diabetes related phenotypes. Given the rapid expansion of human genetic and clinical data, by the completion of Year 1, members of the Accelerating Medicines Partnership expect that the Knowledge Portal will have gained access, aggregated and harmonized 200,000 GWAS, 10,000 exome, and 100,000 exome chip data with corresponding and relevant phenotypic data from the studies introduced below.

The potential of bringing data together in this way has been illustrated by three recent examples. First, a null mutation in APOC3 was found in the Amish population and shown to reduce triglyceride levels and arterial plaque. However, the available information were insufficient to demonstrate whether this effect represented a generic response to loss of APOC3 function, and critically, whether loss of APOC3 function resulted in protection from cardiovascular outcomes.

Recently, exome sequencing and exome chip genotyping have identified four different variants in APOC3 that reduce activity, reduce triglycerides, and collectively protect against cardiovascular events. The combination of these different studies suggests the therapeutic hypothesis that reduction of function of APOC3 might reduce risk of heart attack through a novel mechanism, and that triglycerides might be used a biomarker to monitor therapy.

A second example comes from the zinc transporter SLC30A8 in T2D, in which human genetic data appear to reverse the therapeutic hypothesis as compared to prevailing wisdom based on cell and animal experiments. Five years ago, a common coding variant in the gene SLC30A8 was found in an early GWAS study as a risk factor for T2D, and subsequently shown to be associated with both glucose and insulin levels. (The variant is not associated with adverse events such as cardiac disease or cancer.) The SLC30A8 gene encodes a zinc transporter ZnT8 in the insulin-containing granules of beta cells, and cell and animal studies have shown that knocking out SCL30A8 reduces zinc content with inconsistent effects on glucose homeostasis. Nonetheless, a standing hypothesis is that increasing ZnT8 activity might be beneficial in T2D. Recently, unpublished analysis involving 150,000 samples from multiple populations and many research groups has shown that rare, protein-truncating (loss of function) variants in *SLC30A8* *protect* against T2D in human populations. This result only became clear by harmonizing the data and analysis from many studies, and asking the right question related to therapeutics and T2D.

Thirdly, the value of such an approach is not limited to targets that show a desired relationship between genetic variation and disease – sometimes, the absence of a desired relationship can be highly informative. For example, endothelial lipase (gene name *LIPG*) has been considered a target for pharmacologic inhibition to raise HDL-cholesterol levels and (hopefully) prevent MI. A recent genetic study (again involving >100,000 samples drawn from multiple research groups) showed that while a genetic loss-of-function variant in endothelial lipase did raise HDL-C, it did not have any effect on MI. This led at least one company to kill its endothelial lipase inhibition program – saving (they estimated) tens or more millions of dollars.

In theory, it might seem straightforward to learn from the accumulation of clinical and genetic data. Moreover, it might seem that it will happen on its own, and that companies can simply wait to read the publications that will emerge. In fact, it is highly challenging at present to learn from the data, because it exists in silos (by investigator, by trait, by institution), and because harmonization of technology, analysis, collaborative agreements and IRB approvals requires a focused effort. Each success described above (APOC3, SLC30A8, endothelial lipase) struggled to overcome a similar set of substantial barriers (logistical, regulatory, technical, computational, and statistical), took years to publish, and represent only a tiny fraction of what can be learned by a systematic effort.

Many more datasets exist that could inform target validation, but they simply have not been accessible or assembled in a manner that can supported the needed analyses. Importantly, no previous effort has made the results of all these highly relevant types of analyses freely available so all participants could learn (without collaboration) about targets of potential interest. For a small fraction of the cost of the data that has already been collected, and an even smaller part of the savings that could be realized by avoiding a single phase 3 program that need not be performed, a vast acceleration can be achieved in uncovering, learning, and understanding what experiments of nature can tell us about targets for T2D in man.

Experimental design

Our goal will be to access, aggregate, harmonize, and analyze data on genome sequence and T2D and to make available the results of analyses relating each gene in the genome (and thus fundamental biological processes) to T2D clinical features and outcomes. We will also gather, align, and make available summary data for each gene / variant for related traits (such as glucose, lipids, coronary heart disease, and other macrovascular and microvascular complications of T2D), streamlining the understanding of the full phenotypic consequence of each gene perturbation. The results of all analysis will be available through automated queries so that results can be immediately available to guide target selection and validation. Data, methods, and results will be updated regularly so that the Accelerating Medicines Partnership's T2D Knowledge Portal provides a current view of what can be learned at any time.

Given the number and diversity of genetics studies conducted worldwide, it will be necessary to prioritize for inclusion in the Knowledge Portal those studies that address the following: (1) achieving the minimum number of patients needed to enable robust statistical analysis, (2) enriching the Knowledge Portal to enable impact of ethnic diversity on associations to be investigated, (3) assuring adequate phenotypic data to identify variant / phenotype associations, (4) allowing associations between genetic variants and risk of cardiovascular and kidney disease in T2D to be investigated, (5) inclusion of longitudinal data to assess association between genetic variation and disease trajectory, (6) the capability to callback subjects to assess the physiological/mechanistic impact of certain variations, and (7) the necessary consent parameters to perform the aforementioned analyses.

Achieving these goals requires combining a number of steps, each based on well-established methods, focused on a shared goal of making known the relationships between genetic variation and T2D.

Steps will include:

- Identify / prioritize datasets, obtain permission and access, and gather data
  - Identify high quality datasets on exome and genome sequence in clinical samples that have been deposited in public databases (*e.g.*, dbSNP), and / or that will be shared by cooperating investigators. A core dataset of >20,000 samples (T2D and controls) exists based on the investment in genome sequencing by the GoT2D, T2D-GENES, NHLBI Exome Sequencing Project, and many others. Based on the experience of the DIAGRAM, MAGIC, T2D-GENES and other Consortia (many of which we lead), we believe that much of the data listed above could be contributed or obtained for this purpose.
  - Continuously recruit a sufficient number of robust racial/ethnic minority cohorts to enable subgroup-level analysis and analysis of gene-phenotype interactions, particularly among Asian cohorts as one-third of diabetic patients are in Asia. If after 2 years, the Steering Committee finds the diversity of recruitment unsatisfactory, a plan will be developed to correct for deficits in certain groups by refocusing cohort recruitment strategies.
  - Because cardiovascular disease is a major sequela of T2D and any new medicine for T2D must be demonstrated to not increase CHD risk (and would ideally reduce CHD risk), we will prioritize inclusion of data from large CHD/MI consortia including MIGen, CHARGE, NHLBI Exome Sequencing Project-Early-onset MI, CARDIoGRAM, and C4D.
  - Because of the critical importance of information on kidney disease in T2D, we will actively seek access to datasets (to be specified) and collaborate with investigators that have high quality data on these complications. We particularly note the opportunity for collaboration with ongoing projects such as the IMI SUMMIT project.
  - To illuminate non-coding variants that influence disease (*e.g.*, in GWAS regions) we will identify and aggregate information (and where available raw data) on gene expression quantitative trait loci (eQTLs), for example from the NIH-funded GTEx project.
  - To illuminate metabolic alterations that lie within the pathways linking genetic risk and diabetes risk, and that can serve as biomarkers, we will aim to identify and bring into the Knowledge Portal data on samples characterized by sequencing, metabolite profiling, and outcomes. (list of candidate studies to be provided)
  - For each such dataset, we will obtain needed IRB permissions for analysis in the Knowledge Portal. Because the knowledgebase will focus on providing broad access to results (but not redistribute or provide broad access to individual level data), these regulatory permissions should be straightforward.
  - Wherever possible, we will obtain the raw individual-level data on sequence and phenotype so that harmonization and further analysis can be performed.
  - Where individual level data cannot be obtained, and for GWAS data for T2D and

  - related traits, we will obtain summary statistics of genotype-phenotype relationships from high quality, well powered analyses.
  - The identification, curation, and aggregation of data will continue on an ongoing basis as new sequence datasets become available (from the community or collected by the T2D TV effort as described below in Programs B and C).

- Creation and contribution of new datasets
  - As described in Diabetes Program B, the core dataset described above can and should be augmented by targeted collection of data from populations, phenotypes, and genes of high interest for target validation purposes.
  - Some of these datasets will be collected in the course of research, make their way into dbGAP, or be contributed by the investigators.
  - Another rich opportunity for increased sample size and contribution in-kind might be data obtained from clinical trials (for example, of cardiovascular disease) that included both diabetics and non-diabetics, and that collected phenotypic data at baseline (prior to drug). Even if no data from drug treatment and outcome were shared, the baseline data (genotype and phenotype) alone could be powerfully deployed for target validation in T2D.
- Computational infrastructure for storage, harmonization, variant calling, and quality control
  - Each sequence dataset will have been collected using somewhat different methods for sequencing and for analysis. Thus, harmonization of data and methods will be required to obtain valid estimates of genotype and phenotype.
  - These methods will be optimized, automated, and updated as needed so that they can be run repeatedly as data accumulates, keeping the aggregated data current.
  - Sequence data will be compressed for efficiency and processed through a variant calling layer using consistent approaches to alignment, estimation of error modes, identification of variant sites, and genotype refinement.
  - Summary statistics of genotype (GWAS) data will be harmonized for interpretability.
  - Phenotype data will be harmonized for the relevant traits of interest. For much of the data listed above, this process has already been initiated, and agreed upon methods will be applied uniformly across datasets.
  - Automated comprehensive quality control metrics will be calculated for each dataset and filtered using consistent approaches to arrive at harmonized genotypes and phenotypes.

- Automated annotation and analysis of genotype-phenotype relationships
  - An analysis team will be assembled from among leading T2D statistical geneticists. They will develop a joint analysis plan and contribute methods that will be deployed across the harmonized data. These methods will address confounding due to technical artifacts or population stratification, will enable joint analysis across cohorts and studies, evaluating both common and rare variants individually and together for burden tests in genes and pathways.
  - Annotation methods will be developed and selected for prediction of function consequence of each variant.
  - Methods for annotation and analysis will be automated in an analysis layer so that they can be uniformly performed across the assembled data and updated on a

    regular basis.
  - The analysis team will review and approve each new set of results (as new methods are deployed, and new data added), and iteratively improve methods for maximal power and interpretability.
  - Estimates of association and statistical significance will be calculated for individual variants and burden tests in all genes.
  - Pathway based methods will be selected and deployed for queries related to pathways rather than individual genes.

- Query and visualization layer
  - Automated methods will create results for all genes, but most users will be interested in individual or subsets of genes or pathways.
  - An automated query layer will be developed that allows selection of subsets of tests, datasets, phenotypes, and genes based on the user's interests. The query layer could also be used to control access to particular datasets and results should any elements of the underlying information not be publically available.
  - The Knowledge Portal and project will support two primary levels of query:
    - Original queries by point-of-access users to generate immediate results;
    - Queries developed collaboratively by the partners represented on the Steering Committee.
    - During the early phases of Knowledge Portal development, contributors will have access to the aggregated data and results from a subset of the collaboratively developed queries in a noncompetitive framework. The analysis from those queries will be available for up to six months before they are publicly released via peer-reviewed publication or through other means of public dissemination *e.g.,* publishing on the Knowledge Portal or FNIH website.
  - Supported queries should include:
    - Phenotype-based queries:
      - What are the genes, functional elements, or variants that protect from or contribute to T2D risk, are associated with variability in T2D-related traits, or impact the risk of T2D subjects developing cardiovascular or kidney disease?
      - What are the gene expression associations, or variants associated with T2D risk and T2D-related traits (tissue eQTLs) in relevant tissues (*e.g.*, muscle, liver adipose) from T2D subjects and controls?
    - Gene- or pathway-based queries:
      - What genetic variation (of a specific annotation type: e.g. LoF or GoF) exists within a specific target or pathway? Are these variants associated with T2D risk, quantitative T2D-related traits (*e.g.*, metabolic or lipid), or the risk of T2D subjects developing cardiovascular or kidney disease?
    - Variant-based queries:
      - What are the clinical, biochemical, eQTL, and epigenetic phenotypes associated with this variant?
    - Subset queries:

        - *e.g.*, are results consistent across ancestry groups? Across studies?
    - An automated visualization layer will provide methods for exploring and comparing results across genes, genetic models, and phenotypes, allowing each user to examine specified hypotheses about relationships between gene function and T2D.

We anticipate that these steps will occur in the following sequence:

- **Selection of datasets:** by a joint committee of academic and industry participants
- **Aggregation of data:** by study staff with expertise in data management and informed consent
- **Data storage and variant calling:** by study staff expert in analysis of next-generation sequencing
- **Harmonization of phenotype data:** by study staff guided by a joint committee
- **Development of analysis plan and quality control:** by analytical committee to include academic and industry participants, supported by dedicated study staff
- **Automation of analysis, query, and visualization**: by study staff
- **Learning:** everyone!

A core dataset (as described above) already exists, and joint analyses have been performed in many cases. Thus, no fundamental barriers exist to performing the project. In time, however, we envision that the Knowledge Portal will grow as additional datasets and methods are contributed. Diabetes Program B provides an example of how this foundation could be extended. However, Diabetes Program A can stand alone, and be highly valuable, even if no other data are collected. To the extent that target validation efforts use data from human genetics, Diabetes Program A will provide a better powered, analyzed, and more definitive set of answers on which decisions can be made. Furthermore, it is worth noting that the analytical pipelines described above been individually demonstrated in the course of projects such as the 1000 Genomes Project, DIAGRAM, GoT2D, T2D-GENES, NHLBI Exome Sequencing, and other projects. However, they have not been combined in this way, nor have they been combined with sophisticated software engineering for automation, query, and visualization.

Program B: "Deep genetics" to validate the efficacy and safety of potential therapeutic targets for type 2 diabetes and its complications

Background

Over the past 8 years there has been an explosion in the generation and analysis of large scale data sets that address the relationship between DNA sequence variation and individual risk of type 2 diabetes, and/or related clinical and physiological phenotypes. As of mid-2013, over 70 loci have been shown to be associated with risk of type 2 diabetes to stringent levels of significance; the majority of these identified through genome wide association scans. There has been a similar yield of loci influencing other diabetes-related quantitative traits (including continuous glycemic measures). These discoveries have provided important clues regarding trait architecture (for example, the relative contributions of defects in insulin secretion and action; and the relationship between physiological and pathological variation in glycemia), and insights in mechanisms underlying disease pathogenesis (highlighting enrichment for genes implicating in cell cycle regulation and adipocytokine signaling, for instance). However, there has been only modest progress in turning these genetic associations into detailed mechanistic maps of disease pathogenesis, and thereby identifying promising targets for therapeutic modulation. At very few of these 70 loci, have researchers yet established the causal transcript (i.e. the gene through which the T2D-risk is mediated), the direction of the effect, the mechanism of action, or the full impact on human physiology.

Many factors have contributed to this gulf between locus discovery and biological inference. First, the variants identified by genome wide association scans are common, ancient, of modest effect, and they map overwhelmingly to regulatory, rather than coding, sequence. At most loci, it has not yet been possible to tie down the specific causal variants (due to local patterns of variant correlation), nor to connect the associated regulatory variants to their downstream targets. Second, limited access to human islet, liver, and other metabolic regulatory tissue material has meant that the disease-specific regulatory mechanisms in these key tissues are only now being described, further restricting efforts to define regulatory links between associated variants and nearby transcripts. Third, whilst there has been great progress in the aggregation of summary-level data for discovery purposes, the individual-level data that would support more detailed fine-mapping and functional studies is generally not accessible for combined analysis. As a result, the biological and translational yield from GWAS has, to date at least, been modest.

However, the array-based genome wide association studies which have dominated the human genetics landscape in recent years have focused predominantly on common sequence variants. The availability of more accurate, more affordable sequencing technologies is enabling a shift towards sequence-based discovery efforts, based around whole-genome, whole-exome, and targeted sequencing designs. One of the key features of this shift is the capacity it provides to extend large-scale association discovery efforts to variants with lower allele frequencies (“rare” as opposed to “common” alleles), and thereby to screen for alleles which might have greater functional effects. Since large-effect alleles tend to be under strong selective pressure, they are almost always of relatively recent origin, and therefore present at only low frequency, often only in individuals from a restricted ancestral group.

Rare, large-effect alleles represent particularly valuable accidents of nature which can provide unique insights into the consequences of long-term perturbation of gene function in man. These insights have direct relevance to the need for pharma to understand whether therapeutic modulation of the protein products of those genes (or of other proteins in the same pathways) is likely to be effective and safe. There are a growing number of celebrated examples of the value of such discoveries, including *PCSK9* and coronary disease, and *CCR5* and HIV (see Program A text for further examples). In the T2D realm, in unpublished work, some of us have shown how the identification of rare, protein-truncating variants in the ZnT-8 zinc transporter (encoded by *SLC30A8*) has provided definitive evidence that, in man, loss (rather than gain) of ZnT-8 function is protective against type 2 diabetes, a finding which has obvious implications for existing strategies to define agonistic therapeutics at this compelling target.

The comments above have centered on the pathogenesis of T2D itself. However, much of the morbidity and mortality associated with T2D arises from its complications, both macrovascular (coronary disease, cerebrovascular disease, peripheral arterial disease) and microvascular (nephropathy, retinopathy, neuropathy). Improved glycemic control, allied to available therapeutic options (statins, ACE inhibitors for example), has blunted the impact of some of these complications, but there remains a substantial demand for novel approaches to treatment and prevention. Furthermore, any new therapy for T2D must be shown to be safe with regard to CHD risk (and would ideally reduce risk of MI). Obstacles to drug development include inadequate understanding of the key molecular processes mediating the relationship of T2D to macrovascular disease and the microvascular complications of diabetes, and the paucity of surrogate biomarkers which could be used to stratify risk and characterize progression. The identification and characterization of genetic variants influencing individual risk of these complications has the potential to address both obstacles, but research efforts have lagged behind those of T2D itself, both in terms of the sample sizes deployed, and the numbers of loci for which there is robust evidence of association. The prospects for scientific advance look most favorable for macrovascular complications, where existing large-scale genetic efforts for CAD and MI are in place: these studies already include many diabetic individuals, and attempts are now underway to compare genetic risk profiles in diabetic and non-diabetic subsets of these data, to determine whether there is any interaction with diabetes status. In comparison, the status of human genetics efforts with respect to the microvascular complications of diabetes is considerably less well developed in terms of the existing data and knowledge base, the clinical resources available, and the collaborative infrastructure. However, for diabetic kidney disease, at least, there are a number of well-powered GWAS and sequencing efforts now in place (via the GENIE and SUMMIT consortia for example): over the coming year or two, these should start to provide the evidence base and sample sets that will support analogous efforts within Programs A and B.

The overall strategy of the diabetes project is to provide access to high-quality human genetic data that will allow the evaluation of the efficacy and safety of potential therapeutic targets for diabetes, and its complications, and thereby inform the drug development pipeline. Within Program A, we describe the ways in which we can harness available, and emerging, human genetic data to this end; and in the future opportunity for additional investment in Program C, we explain how detailed physiological and genomic examination of individuals carrying alleles of particular interest can support this objective. Here, in Program B, we focus on the generation of new human genetic and genomic data for targets of particular interest (those for which the existing human genetic and genomic data available through the Knowledge Portal are insufficient to permit a robust “go/no go” decision) through what we have termed “deep” genetics. As we describe below, an initial effort focused on the development of T2D itself, will be extended to provide equivalent insights into target validation for the cardiovascular and diabetic nephropathy complications of diabetes.

Experimental design

In addition to aggregating and converting available data into knowledge (Program A), there will be a need, for many therapeutic targets of interest, to supplement the data currently available through “deep” genetic studies to fill gaps in the knowledge base and thereby to seek to provide actionable information through human validation.

For any given target, we expect Program B to generate data that contributes to answering five key questions:

- Is there evidence from human genetics that lifelong perturbation of that target’s function (or expression) leads to a change with respect to T2D status, diabetes-related quantitative metabolic traits or risk of diabetes complications, which is consistent with the expected outcome of therapeutic modulation?
- Is the desirable therapeutic modulation to be achieved through loss or gain of function?
- Is there evidence from human genetics that lifelong perturbation of the target’s function (or expression) leads to “on target” adverse risk effects that would compromise its value as a therapeutic target?
- Does human genetics or genomics deliver any information with regard to the mechanism of action?
- Can human genetics identify individuals carrying alleles of particular value (most likely rare variants of large effect) that would be of particular interest for consideration for future investment in call back studies described in Program C?

Sequence Target definition

The “genetic [or sequence] targets” of this endeavor will be defined primarily by pharma and will comprise the set of genes encoding potential therapeutic targets for which more intensive human genetics validation is deemed to offer value. Additional genes derived from academia-led genetic discovery efforts could also be added to this list, where these point to entirely novel potential therapeutic targets, or where those discoveries help to define additional targets of interest in cognate pathways.

Because of the focus on large-effect alleles, the vast majority of the sequence to be targeted for study is likely to be derived from protein-coding content. However, it may be desirable in some settings to consider key regulatory sequence for those genes, and also to expand to non-coding RNAs of potential therapeutic import. For a subset of signals where a single variant seems to be driving the association, it may be appropriate, and expedient to use genotyping rather than resequencing approaches.

Given the precompetitive nature of the enterprise, and the economies of scale and cost that can be achieved by combining genetic targets into a single “capture” experiment, we would envisage that the therapeutic target list would be nominated and collated across those pharma companies participating in the T2D component of the Accelerating Medicines Partnership. Interrogation of these genes and their cognate pathways through the Knowledge Portal (within Program A) would provide an initial triage step, with only those deemed in need of additional genetic data submitted for evaluation under Program B. For example, for some genes of interest, low frequency variant data from GWAS and/or imputation already accessible within the Knowledge Portal may provide valuable clues to target validation. The targeted sequencing data from an existing list of genes that is currently being generated across large population cohorts from publicly funded mechanisms is expected to be deposited in the Knowledge Portal in an ongoing, expedited manner to enable this triage step to support focused, non-redundant investments in re-sequencing efforts within large, assembled T2D cohorts. However, such data may not be conclusive, and the opportunity to gather information on a broader range of loss-of-function alleles will be of considerable importance. In addition, for some targets of interest, existing exome sequence and exome array data may not, for technical reasons, have provided adequate coverage of relevant variation, leading to the need for targeted re-sequencing efforts.

Technologies

The additional genetic information will be collected primarily through targeted sequencing, though some signals might be tackled through genotyping of particular variants. The choice in any particular setting will depend on the allele frequency spectrum of the variants of interest.

Strategies and technologies for efficient targeted sequencing on a population-scale (>10,000 subjects) are evolving rapidly. The greatest challenges no longer relate to the generation of raw sequence data, but rather to: (a) processing of many thousands of DNA samples; (b) optimization of capture coverage and efficiency; and (c) mitigation of the high per-sample costs of sequence library construction. The two approaches which, currently, have the best performance in our hands are Illumina's Golden-Gate based TSCA platform, and Agilent's Haloplex. With either of these current technologies, it is possible, at scale, and in individually barcoded samples, to achieve excellent coverage across ~2MB of sequence (equating to the coding content of ~250 genes) for ~$100. This equates to ~4c per exon and is thus already about two orders of magnitude below the costs of Sanger sequencing. However, we are hopeful that, over the course of the project, there will be further reductions in cost and increased capacity for flexibility in terms of sequence target size, and thus the number of genes to be sequenced in this project could be substantially higher, providing greater flexibility in inclusion of specific gene targets.

Samples and cohorts

Targeted analysis of genes of interest will need to be conducted across a wide range of samples using standardized reagents and protocols, with (largely, but perhaps not exclusively) centralized data generation and analysis.

In the first round of analysis, the focus would be on target validation with respect to T2D. The range of samples we would plan to examine would include:

- Large-scale multiethnic T2D case-control sample sets: the obvious starting point here would be the 50K case-control samples (5K cases, 5K controls from each of 5 major ethnic groups) currently being assembled by T2D-GENES for a targeted sequencing study planned for late 2013. This could be complemented by at least ~30K further case-control and cohort samples held by core T2D-GENES investigators;
- Samples from population isolates and consanguineous populations ("an isolate biobank") to enhance prospects for encountering additional "sentinel" alleles, some of which will have risen to higher allele frequencies due to the specific population history. A number of such samples (*e.g.,* Finnish, Ashkenazim) are already included in the ~20K exome sequences collected by the T2D-GENES and GoT2D consortia (to be made available via the Knowledge Portal), and others would be included in the 80K samples referred to above. We would aim to solicit additional samples from suitable populations: amongst those for which T2D case-control samples are already available include isolates of Amish, Native American and European (Dutch, Croatia, Orkneys and others) origin, as well as Pakistani and Arabic samples with high rates of consanguinity. Additional samples to be solicited would include enrichment in South Asian and East Asian T2D case and control isolates.
- Prospective cohorts to enable progression and biomarker studies (for example, the Botnia and

Malmo studies and EPIC/INTERACT studies, where it would make sense to focus on nested case-control or case-cohort subsets;

- Cohorts available for genotype-based recall: several of the samples described above are consented for this, and others are available. For example, in the UK, ~100K individuals are consented for genotype-based recall through the Oxford Biobank, EXTEND, ALSPAC, GO-DARTS and INTERVAL studies; and stored biosamples could be recovered (*e.g.,* for biomarker profiling) in ~500K subjects from the UK Biobank).
- Large clinical trials. An obvious source of samples here would be large CV outcome traits, which will include many subjects with T2D. These will typically have longitudinal samples well characterized for a range of adverse events, as well as disease progression. The emphasis could be on baseline phenotype data and longitudinal data in the placebo control group so as to avoid issues related to outcomes on the drug studied.
- Cohorts with a wide range of phenotypic and outcome data to enable detection of "on-target" adverse effects (eg UK Biobank, Nordic biobanks cohorts, and EMR-based cohorts such as at CHOP or Vanderbilt).
- Inclusion of cohorts with consent for genotype-based recall will be of particular value to enable potential future investment in call back studies in Program C.

These samples would be collectively informative for (a) T2D status; (b) diabetes-related quantitative traits; (c) coronary artery disease and MI; and (d) wider “phenome” data relevant to adverse effects. For most of the above purposes, a combined data set exceeding 100,000 to 150,000 individuals would be feasible and well-powered under a range of realistic genetic models (see below). A large proportion of those samples (>80K) are already being gathered (at Broad) under the auspices of the T2D-GENES and GoT2D consortia. There would be work to do to harmonize clinical and outcome phenotypes to maximize data value. The examples cited in Program A text provide clear illustrations of the value of such broad sampling: the original *PCSK9* finding was made in African-Americans, and an *APOC3* loss of function variant was first found in the Lancaster County Amish, whilst the “sentinel” alleles in *SLC30A8* were detected in a Swedish isolate in Finland, and in individuals from Iceland.

The detection of “on-target” adverse effects relies more on access to large biobanks with copious amounts of phenotype data, and ideally, those which have already been genotyped for many of the variants of interest. The UK Biobank (500K, being genotyped for a combined GWAS/exome array) would be an excellent example. In contrast to the other sample sets described above, we do not envision that a full set of samples from these very large cohorts will be made available specifically for this project: rather that these cohorts be interrogated for variation in genes of interest based on the existing deposited genotypes, or, where necessary, by dedicated genotyping or sequencing efforts.

The data initially generated would be analyzed to address the questions related to T2D, and to target potential for future investment in recall-based studies (Program C) to the most rewarding samples/ethnic groups/isolates.

In the second round of analysis, we would focus on macrovascular complications and diabetic nephropathy. For macrovascular complications there are already large GWAS and sequencing efforts for CAD and MI (*e.g.,* CARDIOGRAM, C4D, ESP) on a scale equivalent to those for T2D (see Program A text), and many of the individuals within these studies are diabetic. There is

already an effort ongoing within CARDIOGRAM and SUMMIT to parse the diabetic contingent of these data and define the genetic determinants of diabetic macrovascular disease (and how these might differ from CAD in the non-diabetic state). We would propose therefore to build on these existing sample sets (as well as reusing some of the samples described above) to support deep genetic studies for proposed therapeutic targets influencing macrovascular disease risk.
The basic concept would be to address the question of genetic sequence differences in the 'target genes' among patients with T2D who developed CAD/MI compared with those with T2D who are older and have not developed CAD/MI. This would address the hypothesis that perturbation in the function of certain genes predisposes to (or protects against) CAD *in the presence of T2D*.
Because ischemic stroke is also a major macrovascular complication of T2D, a parallel experiment would compare cases of ischemic stroke in patients with T2D to control T2D patients without stroke. It is estimated that the CAD/MI experiment would include approximately 40,000 subjects and the stroke experiment would include approximately 20,000 subjects.
The second round of analysis within Program B would also be directed toward diabetic nephropathy as a complication of diabetes. There is substantial interest from pharma in this area, and a sense that, given the poor state of current mechanistic understanding, the clinical benefits to be gained from human genetic studies might be particularly marked. However, as described above (and in Program A text), research into of the human genetics of microvascular diabetes complications currently lacks the "maturity" of that of T2D: the evidence base (in terms of available data for aggregation) is modest, and the range of informative sample sets that could be used for deep genetics or genotype/phenotype studies lags behind. For this experiment, patients with T2D and nephropathy would be compared to those of equivalent duration of T2D without nephropathy. It is estimated that this experiment would include approximately 20,000 subjects.

As noted above, any effort within Program B in this area would involve a distinct set of targets, different sample sets (characterized for the relevant complication phenotypes) and a somewhat different suite of academic and pharma investigators. However, we note that the use of the same conceptual framework and overlapping infrastructure as developed for T2D will encourage efforts to make data aggregation and further genetic and phenotyping studies possible in the middle-term.

We expect that Program B would be able to benefit from ongoing efforts to develop extended sample sets for diabetic kidney disease (for both T1D and T2D). There are two large efforts currently underway in the field: the GENIE consortium of academic investigators and the SUMMIT (pharma-academic) consortium funded under the IMI mechanism in Europe. With the latter due to end in 2014, it might be possible to envisage the TVC to accelerate opportunities to perform equivalent "deep genetics" efforts for diabetic kidney disease targets in the later stages of the project.

Analysis approaches

The sequence and genotype data generated by this effort could already be analyzed through existing pipelines developed for ongoing sequence-based research projects. However, there would be opportunities to develop improved statistical methods (*e.g.*, for studies in population isolates). Data generated would be integrated within, and made available via, the Knowledge Portal.

Sample size estimates for this kind of study will be heavily dependent on the specific characteristics of the alleles we might hope to detect. For the most part, the most severe loss-of-function alleles of greatest interest will be rare, and ethnic-specific, and very large sample sizes indeed, may be required to demonstrate association to established genome-wide levels of significance. However, the biological priors for some of the proposed targets at least, the potential to detect multiple associated alleles in diverse ethnic groups, and the opportunities for detailed follow-up of selected alleles (Program C) will mean that such restrictive significance levels are not essential. (The insights gained from *PCSK9*, *APOC3* and *SLC30A8* variants confirm this view).

For the purposes of illustration, we include here some single variant and gene-based power calculations. The single variant power calculations are based on a total of 19,000 cases and 27,000 controls, assuming an additive model. At the "exome-chip-wide" significance level of $\alpha=4.5\text{x}10^{-6}$, power is high (>99%) for any variant with allele frequency (AF)>1% and RR>1.5. For rarer alleles (AF=0.1%), power remains >99% for RR>3 and >30% for RR=2. At a more stringent genome-wide $\alpha=10^{-9}$, power is high (>90%) for variants with AF>0.5% and RR>1.8. The gene-based power calculations use the C-alpha burden test [Neale et al. 2011] and assume a two-stage design that includes ~10,000 case-control samples exome sequenced in stage 1, and a targeted sequencing follow-up in a further 20,000 case-control samples. We assume a stage 1 threshold (α1m) of .005 (i.e. ~100 transcripts expected under the null) and test models where the combined effect of the risk variants in a given transcript accounts for 0.125%, 0.25%, or 0.5% of liability scale variance in disease risk (assumed prevalence 8%). Power calculations were derived by simulation, based on a model that generates a site frequency spectrum matching that from ~12k European sequenced exomes (as used in exome chip design), and assume a transcript of average size (1.5kb coding sequence). The threshold of declaring significance in the joint two-stage analysis is $\alpha=2.5\text{x}10^{-6}$ (that is, 0.05 corrected for 20,000 genes). We estimate ~75% power to detect, at this joint alpha, a gene explaining 0.25% of the variance rising to >95% for Vg=0.5%.

These illustrative power calculations demonstrate that the kinds of sample sizes we propose here (in programs A and B) should detect a broad swath of functional alleles of interest, particularly when they are segregating across several different sample sets.

Further components

Though the focus in the narrative above is on coding variation, we expect that there will be many naturally occurring variants of interest which map to non-coding variation. Efforts to validate therapeutic targets will be assisted by our capacity to link non-coding variants to regulatory annotations and thereby to the transcripts whose expression they regulate. We see substantial merit therefore in the aggregation of regulatory annotation data for tissues of particular interest to diabetes pathogenesis (islet, muscle, fat, liver) and in investing in the generation of additional data (through RNA-Seq, ChiP-Seq etc) where the existing data sets are inadequate. These studies will allow us to link non-coding variants to their "causal" transcripts through cis- and trans-eQTL mapping, and through the characterization of enhancer-transcript links (as recently demonstrated by the ENCODE project). Indicative budgets for this are provided in section 5.

We anticipate that the proposed program will occur in the following sequence:

The following will occur in parallel during year 1 - 2015:

- **Sequence target definition:** by pharma and academics jointly, benefiting from the results available from the Knowledge Portal and emerging data from current publicly funded T2D resequencing efforts.
- **Aggregation of samples;** by academics principally, building on existing consortia. Prioritisation and selection of samples guided by pharma recommendations;
- **Assay validation:** by academics principally, already in progress.

Thereafter, in year 2 - 2016:

- **Genetic data generation:** by academics, or possibly by biotech/service companies.
- **Data analysis:** by academics with support from pharma as desired.
- **Deposition:** within the Knowledge Portal
- **Interpretation:** by academics and pharma jointly.

Regular, periodic cycles of target definition, deep sequencing and analyses would be envisioned throughout the duration of the project, as feasible. In the budget below, we envisage that a total of two such cycles would be possible during the project. The latter cycle would focus on macrovascular (MACE and stroke events) and diabetic nephropathy complications of diabetes.

It is worth noting that many of the targets selected for deep genetics in Program B, will be those for which encouraging preliminary results have been seen in the analysis of existing data via the Knowledge Portal. Complementary targeted sequencing data generated in the public domain external to the AMP will be incorporated in regular updates to the Knowledge Portal. Additional data will be generated by Program C to confirm and extend those previous findings. Furthermore, this study could be run immediately with existing pipelines, such as the Broad/TSCA pipeline that T2D-GENES will use in late 2013 to perform targeted sequencing of ~250 genes in 50K samples. Data from this effort is expected to be integrated within the Knowledge Portal. This pipeline has already generated proof-of-principle data as part of an MI targeted sequencing study and is undergoing iterative improvement. This TVC project, if funded, would benefit from the experience gained in these previous projects. We would hope that later rounds of target validation in Program B will benefit from the substantial technical improvements in targeted sequencing motivated in the years ahead by strong research and clinical demands.

Experimental design

Programs A and B will generate and collate into a knowledge base a very large volume of previously collected and new sequence data from large numbers of T2D subjects from diverse populations, including some population isolates and consanguineous populations. Functional consequences of some of the variants, e.g., nonsense and frameshift mutations, may be obvious; while the functional consequences of other variants, e.g., missense, synonymous, UTR and intron/extragenic mutations, may be less so. Given the expense and effort required for call-back phenotyping, every reasonable effort should be made to demonstrate that a potential variant of interest is indeed a LoF/GoF variant. Several complementary approaches will be used to prioritize variants of interest including:

- Use of bioinformatics to predict consequences on protein structure and function. An emphasis will be placed on nonsense, frameshift, and splice-site mutations, but missense variants will also be analyzed.
- *In vitro* models to express and compare functional activity of variant versus wild-type protein, particularly in the case of missense variants.
- Evidence for association with T2D or related traits if the variant is common enough for robust statistical analysis (which may be enhanced by genotyping of additional samples from well-phenotyped cohorts)
- Association-based analysis for the aggregate of several rare variants in a given gene.
- Co-segregation analysis in the family(ies) harboring a rare variant of potential interest.

We anticipate that Program A (development of a knowledge base of existing data sets) and Program B (targeted sequencing in large numbers of phenotyped individuals) may uncover previously unappreciated LoF/GoF mutations worthy of call-back phenotyping. These variants will be fast-tracked for deeper phenotyping and target validation. Thus, it is expected that call-back genotype-targeted phenotyping, as proposed in Program C will unveil new biology from nascent discoveries in existing datasets as a result of Project A as well as from new datasets as a result of Program B. We do not know of a similar systematic approach to target validation for T2D through multi-national collaboration using human genomics approaches.

Program C will integrate seamlessly with Programs A and B to efficiently identify adequate numbers of subjects with high-priority LoF/GoF variants and call-back sufficient numbers of these individuals and matched controls for deep hypothesis-driven phenotyping to discern underlying mechanisms and pathways and to validate novel drug targets. DNA collections already available from research subjects, who agreed in the Informed Consent to be recontacted for potential future studies, will be most efficient for call-back studies. In collaboration with Program A, we will catalog in advance all known DNA collections amenable to call-back, including general population collections and population isolates. Call-backs from populations in which the variant of interest is adequately common can be performed as a new study, ie., in the absence of permission to re-contact research subjects, but at increased effort, cost, and time. This will only be done as a last resort. By contrast, if a rare variant is enriched or present only in a population isolate (as determined by genotyping of the population isolate biobank (see below)), there will be no choice but to approach this population for additional phenotyping. Recruiting family members may be an effective approach to increase sample size for call-back studies as well as to possibly identify homozygotes, especially in consanguineous populations. Rare variants in one population may be more common in other populations, especially population isolates due to drift. Indeed, some LoF/GoF variants may be private in a given population isolate justifying primary sequencing of T2D subjects from multiplex pedigrees of population isolates (Project B). Variant(s) providing the greatest evidence for being true LoF/GoF variants will be targeted for call-back phenotyping.

Identification of subjects with LoF/GoF variants of interest for call back:

- For common variants in a given "general" population, if not already known through sequencing (Program B), we will genotype the LoF/GoF variant(s) of interest in large numbers of DNA samples from the population in which it was found. For variants of apparent sufficiently large effect, all three genotypes will be studied since the heterozygotes may provide insights into "dosage" effects that may be relevant to the biology and provide clues to the extent to which the target needs to be inhibited or activated.

- For rare variants, we will attempt to identify a population in which the frequency of the variant is increased. This approach will be greatly facilitated by constructing a population isolate registry and biobank, *e.g.,* 100 or so samples from each population isolate, from as many population isolates and consanguineous populations as possible. Development of a population isolate biobank will be invaluable not only for T2D target validation, but also for the study of diabetes-related complications as well as for other diseases of interest to the broader TVC effort. Some of these populations (if phenotyped for T2D) will be part of the Project B sequencing effort and the frequency of rare variants of potential interest immediate known, while the frequencies of rare variants of potential interest from other population isolates can be ready known by follow-up genotyping of the population isolate biobank. Populations identified as having a high frequency of the variant of interest will be pursued further with genotyping of larger numbers of samples to identify additional mutation carriers and if possible, homozygotes. Family data if available will be used to identify additional rare mutation carriers, and even homozygotes. This approach has been used highly effectively in several population isolates, for example in the Lancaster County Old Order Amish.
- Establishment of a biobank requires a significant ongoing investment. Samples collected and stored as a result of Program C will be extremely valuable and limited in quantity. Establishment of a de novo biobank for this purpose may not be cost effective. NIH staff will need to explore whether a T2D deep phenotyping biobank can be attached to an extant effort to minimize cost.

Hypothesis-driven phenotyping of subjects with LoF/GoF variants of interest:

- In general, for LoF/GoF variants that influence T2D risk, it will be desirable to study mutation carriers who are non-diabetic, because the diabetic state (and medications) may confound the physiology.
- Some routine phenotyping to be performed locally, eg., BMI, fasting bloods, OGTT, will be efficient in getting initial insights into biology and will aid in selecting those individuals in whom deeper phenotyping will be performed.
- The specifics of deeper phenotyping will to a certain extent be based on the known biology of the genes in which the LoF/GoF variant(s) of interest exist, but will likely include a subset of the following:
  - Measurements of body size and composition (egs., BMI, DXA, MRI)
  - Glucose tolerance/insulin sensitivity/insulin secretion (egs., OGTT, FSIVGTT, clamp with tracers)
  - Energy expenditure (e.g., indirect calorimetry, doubly labeled water; BAT assessment by PET)
  - Lipid homeostasis (e.g., lipid profile, subparticle analysis, tracer/turnover studies)
  - Liver steatosis (MRI)
  - *In vivo* substrate metabolism (e.g., tracer studies, NMR spectroscopy)
  - *In vitro* metabolic studies (e.g., functional studies from muscle and fat biopsies)
  - Tissue/biopsy molecular studies (egs., RNAseq; CHIPseq, proteomics)
  - Other serum biomarker and 'omic' studies (inflammatory markers, proteomics metabolomics)

    - iPS cells to assess functional consequences in otherwise inaccessible tissues (beta cells, liver, brain)
    - Phenotyping for on-target adverse side effects (CVD, liver, kidney, brain, musculoskeletal, cancer)
    - Diabetes-related complications (heart, kidney, eye, nerve)
- Depending on the geographic location of the study subjects and the ability of local investigators to perform the relevant phenotypic assessments, these studies may be performed locally, or research subjects transported to designated centers with expertise in the aforementioned human phenotyping methodologies.
- The combination of in-depth phenotyping and the desirability to collect samples for a biorepository, will impose a significant burden on subjects. This will likely make recruitment for such studies difficult; this will be especially true for individuals without T2D (see section on subject selection).
- Variants which implicate novel pathways could present a challenge if interrogation of these pathways would be difficult using existing bio sampling methods,

The number of subjects that need to be phenotyped is dependent on the magnitude of the effect of the variant of interest on the phenotype. We will prioritize LoF/GoF variants that, appear to have a large effect; however, we will also consider more subtle but important phenotypic manifestations of LoF/GoF variants since they may uncover compensatory mechanisms and pathways that themselves may unveil new targets. Given the aforementioned considerations, the ballpark of subjects to undergo deep phenotyping for a given variant will be 10-25 per genotype. This number was chosen because it is adequate to demonstrate (or rule out) a large effect.

Program C will require a multi-national collaborative effort. Logistics will be challenging, but doable. It will be important to pay special attention to regulatory issues surrounding creation of the isolates registry and biobank and call-back protocols. It will be essential to build a culture of open collaboration among investigators both inside and outside the TVC based upon trust, transparency and mutual benefit. Luckily, the T2D community is extremely collaborative and several of us in the T2D TVC have experience building such multi-national consortia.

Each chosen LoF or GoF will represent a major investment in subject selection, recruitment efforts, clinical time, recruitment of appropriate control subjects, and biorepository cost. It is not clear at this time if the benefits of these studies will balance the cost to industry partners.

We anticipate that the proposed program will occur in the following sequence:

Development of a population isolate and consanguineous population registry and biobank:

- As discussed above in Program B, a registry of population isolates and consanguineous populations will be developed. This will be performed by an extensive literature search and personal knowledge and contacts by TVC investigators.
- The PI’s of population isolates will be approached for participation by personal contact or email.
- An initial questionnaire will be used to obtain data regarding the population characteristics, availability of DNA and willingness to contribute de-identified samples to the biobank, availability of phenotype data related to T2D, and permission to re-contact research subjects.

- With appropriate IRB approvals, samples will be shipped to a central location for construction of the population isolates biobank.

Hypothesis-driven phenotyping of subjects with LoF/GoF variants of interest:

- Common LoF/GoF variants in the general population are expected to be rare, but if/when identified, recruitment for call-back phenotyping will be accomplished in these populations. For a rare LoF/GoF variant that is worthy of follow-up, the frequency of the variant, if not already known, will be genotyped in samples from the population isolate biobank.
- The population with a high frequency of the variant of interest and most assessable will be targeted for call-back phenotyping. With the PI of the population, logistic details will be worked out for recruitment of the requisite number of subjects of each genotype and initial local phenotypic assessment.
- Depending on call-back phenotypes of interest and the ability of local investigators to perform the studies, phenotyping will either be performed locally or at a center in which the appropriate clinical research expertise exists. The latter may require transportation of research subjects and other costs. Alternatively, it may be possible for clinical investigators with the appropriate expertise to assist local investigators to set up and perform the aforementioned phenotyping locally.
-

Section III: Project Management

A. Steering Committee Oversight

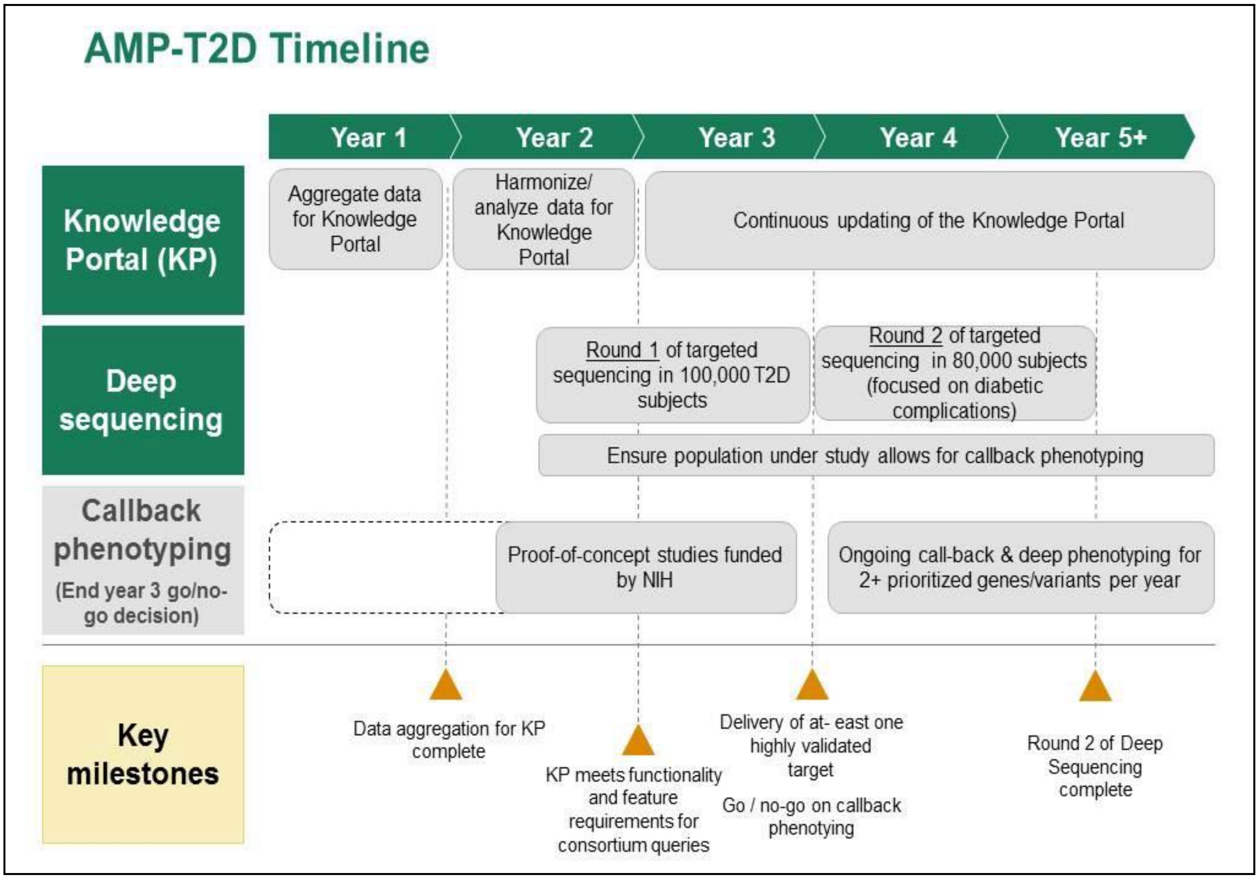
The Type 2 Diabetes Accelerating Medicines Partnership will be governed through a Joint Steering Committee, comprised of members from participating companies, government, academia, and non-profit organizations. The T2D Steering Committee for AMP will operate under the direction of the overall AMP Executive Committee (EC), comprised of 3-4 leaders each from industry and NIH, as well as a representative each from FDA, academia, and the patient advocacy sector. The EC is in turn advised by an Extended Executive Committee comprised of R&D heads of companies involved in the partnership. The T2D Steering Committee is responsible for defining the research agenda and project plan, for review of ongoing projects, and for the detailed assessment of milestones. The project plans are submitted by the Steering Committee to the EC for review and approval. The EC will also review the assessment of milestones and any revision to the project plan that results from a “No-go” assessment that some element of the current plan is not feasible.

B. Governance

The T2D Steering Committee operates under the direction of the Core Executive Committee for the partnership, which is advised by the Extended Executive Committee. The T2D Steering Committee is responsible for defining the research agenda and project plan, for review of ongoing projects, and for the assessment of milestones. The project plans are submitted by the Steering Committee to the Executive Committee for review and approval. The Executive Committee will also review the assessment of milestones and any revision to the project plan that results from a “no-go” assessment that some element of the current plan is not feasible.

Section IV: Timeline, Milestones and Deliverables

A. Timeline and Milestones



AMP-T2D Timeline

Year 1
Year 2
Year 3
Year 4
Year 5+

Knowledge Portal (KP)
Aggregate data for Knowledge Portal
Harmonize/ analyze data for Knowledge Portal
Continuous updating of the Knowledge Portal

Deep sequencing
Round 1 of targeted sequencing in 100,000 T2D subjects
Round 2 of targeted sequencing in 80,000 subjects (focused on diabetic complications)
Ensure population under study allows for callback phenotyping

Callback phenotyping (End year 3 go/no-go decision)
Proof-of-concept studies funded by NIH
Ongoing call-back & deep phenotyping for 2+ prioritized genes/variants per year

Key milestones
Data aggregation for KP complete
KP meets functionality and feature requirements for consortium queries
Delivery of at- east one highly validated target
Go / no-go on callback phenotyping
Round 2 of Deep Sequencing complete

We envision this integrated project involving Programs A and B as a five-year project, in which Program A will begin in the summer of 2014 with the first part of that year used to create a Request for Proposals at FNIH in parallel with the Request for Applications that will be used at NIH, to jointly select the institution at which the knowledge portal will be housed. For program B, starting in the summer of 2014, an inventory of the needed samples and contracts, associated with their transfer will be put in place to allow the launch of Program B in 2015. From that point forward, the two programs will run in parallel until their end in mid-2019. The common oversight Steering Committee for the two programs will ensure that substantial scientific and logistical interaction occurs between them.

Provided below is a list of the specific milestones agreed to by the Steering Committee for inclusion in this project plan:

Knowledge Portal:

1. Selection and complete aggregation of datasets
2. Data storage and variant calling defined
3. Meeting minimum threshold of individuals and cohorts included in the database

4. Collecting a sufficient number of robust racial/ethnic minority cohorts to enable subgroup-level analysis and analysis of gene-phenotype interactions, particularly among Asian cohorts as one-third of diabetic patients are in Asia.

Note that investigators contributing their data sets will be encouraged to share or publish quickly previous analyses done with their data sets in order to contribute to the data integration efforts of the T2D Accelerating Medicines Partnership.

Year 2

1. Harmonization of phenotype data
2. Development of analysis plan and quality control
3. Automation of analysis, query, and visualization
4. Meeting functionality and feature requirements necessary for queries: Phenotype, gene, and variant based
5. Target ID through KP analysis (includes sequencing data)

Years 3-5

1. New target hypotheses generated based on data generated from initial KP results
2. Continuous updating and curation of the Knowledge Portal
3. Expansion of focus from T2D to CAD, other macrovascular disease, and microvascular complications.

Deep Sequencing

Year 1

1. Prioritization and selection of samples for targeted sequencing
2. Aggregation of DNA samples for targeted sequencing
3. Assays validated for targeted sequencing

Year 2

1. Complete first round of targeted sequencing in 100,000 subjects (end 2016)
2. Development of a population isolates biobank
3. Data analysis of targeted sequencing data
4. Deposition and integration of targeted sequencing data within the Knowledge Portal
5. Interpretation of targeted sequencing data

Years 3-4

1. Regular, periodic cycles of target definition, deep sequencing and analyses throughout the duration of the project
2. Continuous updating of the Knowledge Portal with Deep Sequencing data
3. Expansion of focus from T2D to other macrovascular disease, and microvascular complications in 80,000 subjects by end of 2018
4.

Callback phenotyping

Years 1-2 NIH will fund proof of concept studies for callback phenotyping efforts (separately from funding for AMP effort)

1. Deep sequencing efforts will allow for future callback phenotyping

Year 3+ – 2016+

1. Steering Committee will make a go / no-go decision to fund additional callback phenotyping studies based on results of NIH POC study and other partnership efforts

B. Go-No-Go Decisions

1. Meeting the minimum threshold of individuals and cohorts, meeting the criteria as defined in in Section 2 above, such as a level of ethnic diversity, longitudinal follow-up, etc., by the end of year 1 for Program A. A minimum number of 200,000 GWAS, 100,000 exome chips, and 10,000 exomes incorporated into the portal within the first year of project funding.

2. Meeting functionality and feature requirements necessary for queries submitted to the Knowledge Portal within 18 months of project funding.

3. Delivering analyses from Steering Committee-specified research queries for the Knowledge Portal by the end of year 1 (based on datasets available at the time). The Steering Committee members would have immediate access to the results. The partnership would make results publicly available following QC/QA and publication requirements (exclusive data access not to exceed 6 months).

The successful attainment of these criteria will be adjudicated by the T2D Steering Committee. The findings of the T2D Steering Committee will be binding on participants; in other words, if the committee determines that the "go" criteria have been met, individual members will not have an opt out option.

C. Key Deliverables

This partnership aims to determine whether, for gene targets of interest, human genetic data can inform potential drug efficacy and safety. Target-specific validation reports will be generated including analytical summaries of genetic variants identified through targeted sequencing, prediction of variant impact on function, statistical strength of relationships of those variants to T2D disease risk, quantitative metabolic traits, and cardiovascular risk, and for those subjected to deep phenotyping the results and conclusions of such studies.