Developing a model for collaborative science: a mid-term perspective on the AMP T2D Partnership

by Maria C. Costanzo, Manager of Content and Community, Type 2 Diabetes Knowledge Portal

In 2011, Dr. Francis Collins, Director of the National Institutes of Health (NIH), met with leaders in biomedical research to discuss a frustrating problem. Continual improvements in molecular biological and genomic techniques were generating an avalanche of data relevant to complex diseases, yet the translation of these data into insights about disease mechanisms and drug targets was unacceptably slow. It was clear that an entirely new paradigm for collaborative research would be needed to speed up the extraction of knowledge from data.

The result of these discussions was the creation of the Accelerating Medicines Partnership (AMP), one branch of which focuses on type 2 diabetes (T2D)—a life-threatening disease that affects hundreds of millions of people worldwide, whose incidence is growing, and whose progression cannot yet be effectively stopped or reversed. AMP T2D, a five-year project, includes the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK); the pharmaceutical companies Janssen Pharmaceuticals, Eli Lilly and Company, Merck, Pfizer, and Sanofi; the University of Michigan; the University of Oxford; the Broad Institute; and other researchers around the globe. The Foundation for the National Institutes of Health (FNIH) also provides funding and coordination for the project.

Drawing on the strengths of both academia and industry, this public-private partnership brings together all stakeholders in a pre-competitive space to share data and combine resources, with the goal of validating new drug targets faster. Now in Spring 2018, roughly mid-way through the funding period, it is evident that this collaboration has resulted in remarkable progress on both scientific and collaborative fronts.

Genetic association data: the foundation of AMP T2D

Genetic association studies interrogate the genomes of individuals at millions of specific genomic positions to discover sequence variants that are correlated with the incidence of disease. From the outset, AMP T2D aimed to support the generation of unprecedented amounts of new genome-wide association study (GWAS), exome sequencing, and whole-genome sequencing data within the project as well as their aggregation with all relevant publicly available data. Originally, 5 sites were funded by the NIDDK to generate new data and deposit them into the AMP T2D Data Coordinating Center (DCC) at the Broad Institute. As the project evolved, another site was funded by the NIDDK and 8 more sites were funded by the FNIH. Additionally, an Opportunity Pool of funds from the NIDDK was created, allowing the AMP T2D Steering Committee to award smaller grants for complementary research projects in a flexible, science-driven manner. Currently 10 Opportunity Pool projects are in progress, and more awards will be given in the future.

Not only has the number of genetic association studies increased since the inception of AMP T2D, but also the number of samples surveyed in each has grown dramatically, from typically under 100,000 to approaching 1 million today. The increased statistical power conferred by these large sample sizes has led to a huge increase in the number

of loci found to be significantly associated with T2D, from about 70 at the start of the project to nearly 430.

Improvements in genomic technologies in the past few years have allowed AMP T2D collaborators to generate increasing amounts of sequencing data, which make it possible to comprehensively interrogate all alleles and to uncover rare variation. At the project's start, T2D associations with exome sequences (covering the protein-coding regions of the genome) were available for about 13,000 samples, and no whole-genome sequencing studies had been published. Now, more than 2,600 whole genomes are available, and analysis of a set of 50,000 exomes—the largest disease-specific aggregation of exome sequencing data to date—is nearly complete. Importantly, many of the associations that have been newly discovered in sequencing studies involve relatively rare variants that affect protein-coding regions. It is often more straightforward to develop hypotheses about the impact of such variants than it is for variants outside of coding regions.

As the AMP T2D partnership has grown in prominence in the diabetes field, the DCC has been approached by investigators outside the project who want to contribute their data in order to aggregate and display them in the context of AMP T2D data. In early 2017, researchers in the 70kforT2D project, which found novel T2D associations by re-analyzing existing GWAS data, offered their results for integration into the DCC and display in the Type 2 Diabetes Knowledge Portal (T2DKP; see below) before publication.

70kforT2D GWAS was first pre-publication dataset to be added to the T2DKP from outside the AMP T2D partnership, and it was particularly appropriate that these scientists, whose results illustrate the value of data sharing, themselves chose to freely share their results. Incorporation of datasets into the AMP T2D DCC and T2DKP offers investigators the chance to take advantage of the expertise of the AMP DCC analysis team, apply cutting-edge analysis tools to their data, and display their results broadly to the T2D research community in the context of multiple datasets. The AMP T2D DCC is open to incorporating T2D-relevant datasets from all investigators (find details on contributing data here).

In addition to the datasets generated by AMP T2D partners and other T2D researchers, which focus on associations with T2D, glycemic measures, and T2D complications, the AMP T2D DCC also collects publicly available genetic association datasets for traits relevant to T2D, such as anthropometric measures, blood pressure and lipid levels, and heart and kidney disease.

Orthogonal data types to help identify and prioritize causal variants and genes

Finding genetic variants that are associated with T2D risk is critically important to understanding the genetics of T2D, but it is only a first step. The most significantly associated variant in a genomic region may not be the causal variant that is responsible for altered T2D risk. Researchers perform fine mapping to analyze genetic associations in specific regions of the genome and generate credible sets—that is, sets of variants that are predicted to include the causal variant. Mid-way through the AMP T2D funding

period, emphasis among the data-generating partners is beginning to shift from simply generating association data to performing fine mapping and credible set analysis.

But even after predicting which sequence variations are responsible for altered risk, finding clues about how they affect risk requires integration with additional data types. Information about the functional importance of the genomic region where a variant is located—its relevance to gene expression, protein function, networks and pathways, metabolite levels, and more, all determined on a tissue-specific basis—can help prioritize genes and pathways for in-depth experimental investigation. These kinds of research were built into AMP T2D from the beginning, and as the importance of these data types became even clearer, several Opportunity Pool awards were given to projects focusing on complementary data types that shed light on the significance of genetic associations.

Several of these projects focus on generating tissue-specific epigenomic data: histone modifications, DNA methylation, chromatin conformation, transcription factor binding, 3-dimensional chromosome structure, and other data types. Epigenomic data can provide important clues about the mechanisms by which sequence variation affects T2D risk, particularly for variants that lie outside of protein-coding regions. For example, if a risk-associated variant is seen to disrupt a transcription factor binding site, this would support the hypothesis that the transcription factor and its target genes are relevant to T2D.

To make these data accessible to researchers, one Opportunity Pool award supports the creation of T2DREAM, the Type 2 Diabetes REgulatory Annotation Map, which collects and displays epigenomic datasets relevant to T2D. In the near future, these data will be fully integrated with genetic association data in the Type 2 Diabetes Knowledge Portal (see below).

Other Opportunity Pool projects are concerned with processes downstream of gene expression. Discovering interactions between proteins implicated in T2D risk, for example, could help to uncover all of the players in pathways important for the development of T2D, increasing the number of potential drug targets. Determining the effects of variants on the levels of key metabolites can illuminate the metabolic pathways that change during the development of T2D.

In addition to generating all of these orthogonal data types, AMP T2D partners are developing algorithms and using machine learning to classify and prioritize variants on the basis of the functional annotations that accompany them. Finally, other Opportunity Pool projects will use model organisms to test and validate drug targets that are suggested by these analyses.

Tools and methods to speed analysis and interpretation

At the inception of AMP T2D it was also clear that the development of new methods and tools would need to accompany the generation of data, and support for these activities was built into the program. One major technical effort has addressed an obstacle to global data aggregation: because of institutional and national privacy regulations, some datasets may not leave their site of origin to be aggregated with other datasets at the

AMP T2D DCC. A group at the European Bioinformatics Institute has built a technical replicate of the DCC and knowledgebase, such that data stored there are equally as accessible for browsing, searching, and interactive analysis as are the data stored at the AMP T2D DCC at the Broad Institute. This federation mechanism allows global data accessibility even when data aggregation is not permitted.

Other efforts supported by AMP T2D are aimed at improving the speed and efficiency at which data can be taken in and analyzed. In one project, a data intake system is being developed that will streamline the process for both data submitters and for the DCC team, and will be applicable to data submission both at the Broad DCC and at other federated sites. Another project has created a software pipeline, LoamStream, that will largely automate quality control and association analysis of incoming data. Currently, LoamStream is in use for quality control of genotype data, and this has already greatly reduced the time required to process new datasets. Future work will extend the pipeline to association analysis and will also allow it to take in sequence data as well as genotype data.

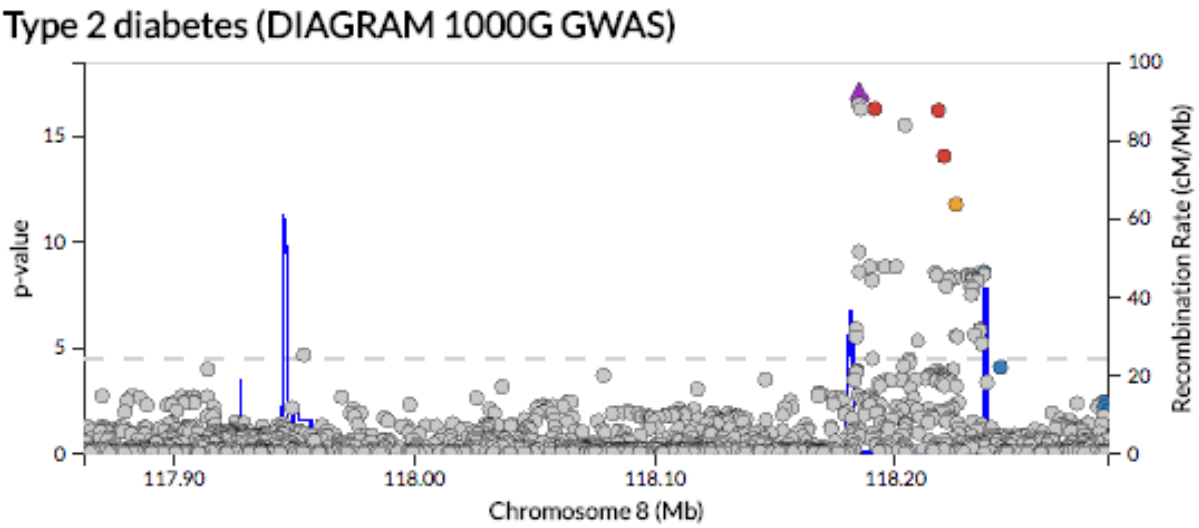
A genetic association of a variant with T2D gains credibility if multiple independent studies replicate the association. Thus, it is important for researchers to be able to evaluate the weight of available evidence. But currently this is difficult to assess from the association datasets in the AMP T2D DCC, because many are based on overlapping sets of subjects. AMP T2D partners at the University of Michigan and University of Oxford are working on a method to take these overlaps into account and synthesize associations from multiple datasets into a “bottom-line” significance for association of a variant with T2D, which will aid in prioritizing variants for future work.

Multiple AMP T2D projects for analysis, interpretation, and custom interactive analysis of variant-phenotype associations are ongoing at the Universities of Michigan, Chicago, and Oxford, Vanderbilt University, and the Broad Institute. These projects are aimed at facilitating, in various ways, the path from variant associations to functional knowledge, and all have been or will be integrated into the T2D Knowledge Portal (see below).

Hail software offers a pipeline that speeds up the analysis of huge genomic datasets, while the gnomAD resource aggregates and harmonizes exome and genome sequences to provide a catalog of genetic diversity, in more than 100,000 humans, that aids in interpretation of variant associations with disease. A tool under development in the gnomAD project will display the effects of variants on protein structures as another way to deduce their potential impact.

Other analysis modules include gene-based association methods for using expression data to predict genes that may impact a phenotype (PrediXcan and MetaXcan), and a phenome-wide association study (PheWAS) method for visualization of the associations of a variant across multiple phenotypes, which is a crucial consideration during drug development.

The interactive visualization tool LocusZoom will integrate many of these methods to display variant associations and credible sets, epigenomic and functional annotations, and phenotype associations across a genomic region as well as offering custom association analysis.



Example LocusZoom plot

AMP T2D Knowledge Portal: democratizing T2D genetic results for researchers world-wide

AMP T2D was founded on the idea that in order to truly accelerate progress, genomic information must be freely accessible to all scientists and presented in a way that is understandable by a broad range of researchers working on T2D biology, not only by human geneticists and bioinformaticians with special computational skills. So the roadmap for the project included not only data generation and analysis, but also the production of a publicly available web resource that would integrate data types, interpret the evidence, and present of all these results.

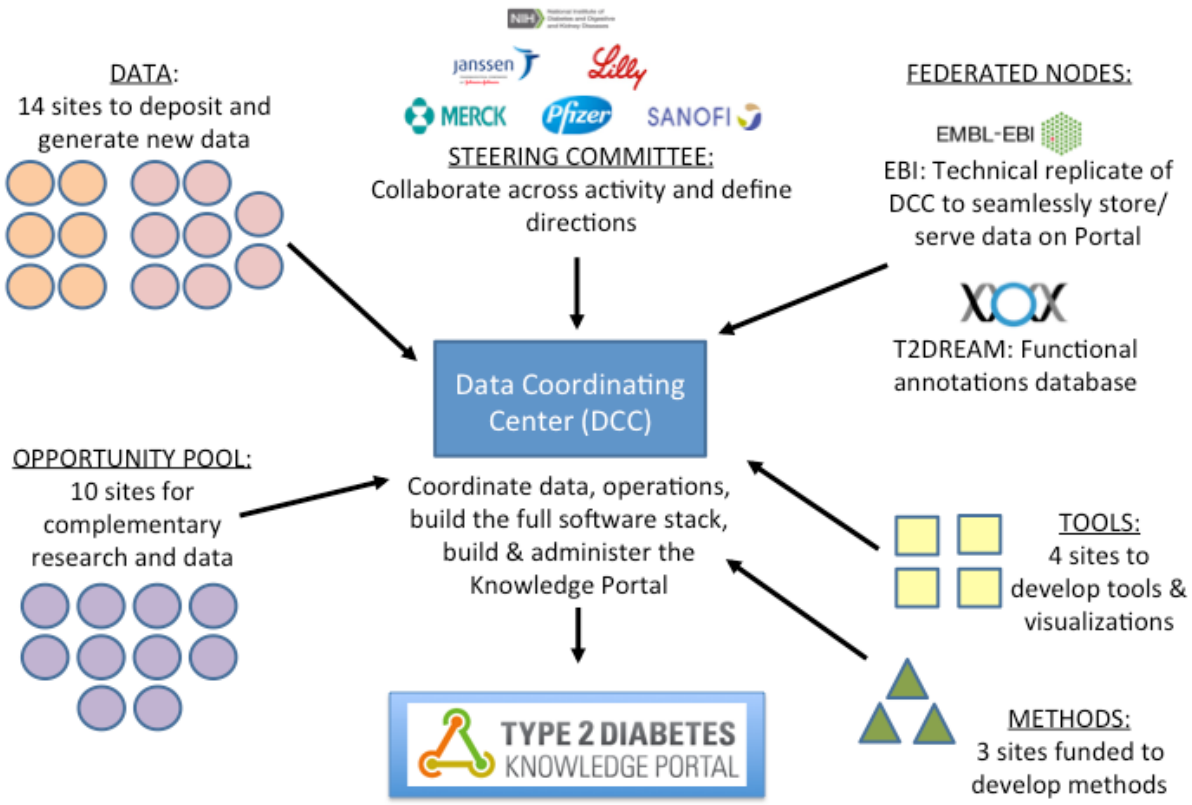
While it is under continuous development, mid-way through the initial funding period the T2D Knowledge Portal (T2DKP) is already a well-established resource. Other web resources collect genetic association data, but the T2DKP is unusual in providing harmonized datasets to which a consistent analysis pipeline has been applied. Rather than simply cataloging datasets, it offers distilled and synthesized results along with their interpretation, to guide more detailed exploration of the evidence. And, unlike any other extant resource, it offers researchers the ability to perform interactive queries on protected individual-level data.

The Gene page of the T2DKP (see an example) illustrates the presentation of immediately understandable summary information along with the opportunity to drill down to the details. An algorithm considers the associations of all variants across a gene, for all phenotypes and in all datasets aggregated at the DCC, and calculates from them a “traffic light” signal for the gene: green to indicate that there is a significant association for at least one phenotype; yellow to indicate suggestive, if not highly significant, associations; and red to indicate that there is no evidence for association for any of the phenotypes considered in the T2DKP. Below this, tables and graphics invite users to explore all variants across the gene, their impacts on the encoded protein, and their associations, as well as their positions relative to epigenomic marks across the region in multiple tissues.

The T2DKP currently offers the ability to run custom, interactive association analyses using two different tools. In the LocusZoom visualization, users may choose one or more variants as covariates before performing association analysis. The Genetic Association Interactive Tool (GAIT) for single variant associations, which also powers the custom burden test for gene-level associations, is even more versatile, presenting the distributions of different characteristics of the sample set (age, sex, BMI, glycemic measures, blood lipid levels, and many more) and allowing users to filter the set by multiple criteria and to choose custom covariates before performing association analysis. Both of these tools allow analytical access to the individual-level data, whether housed at the Broad DCC or at the EBI federated node, in a secure environment so that data privacy is always protected.

T2DKP home page

Evolution of a collaborative environment



AMP T2D organization

The AMP T2D partnership is a multifaceted project (illustrated above) that embraces several aspects of basic research and combines them with building a product, the T2DKP. In connecting scientists both within and outside of consortia, in academia and in industry, working on genetic associations or functional studies, it is becoming the nexus of the T2D genetics community. Researchers are finding the T2DKP helpful for accessing even their own results and for viewing them in the context of multiple phenotypic associations and other complementary data types. Pharmaceutical partners are finding help via the Target Prioritization project, in which the tools and methods developed within AMP T2D are being used to prioritize a list of genes of mutual interest for further investigation.

Perhaps most importantly, AMP T2D has made researchers—both within and outside of the project—aware of the value of sharing data for representation in the context of all other relevant data. Only by compiling and interpreting all available information will we be able to make the best hypotheses about genes and pathways that are possible drug targets and prioritize them for in-depth functional investigation.

AMP T2D and beyond

In the remainder of the initial AMP T2D funding period, we expect continued progress in each of the areas discussed above. The data intake and analysis pipelines will be improved, and new data will be incorporated at an increasing pace—including data from the UK Biobank, which has generated association results for 500,000 genotyped subjects and more than 2,500 traits. Associations will be added for many more phenotypes related to T2D, including diabetic complications and longitudinal phenotype data that connect the development of various traits to the timeline of incident T2D. Much more T2D-relevant epigenomic data will be available for query as well as for browsing, via dynamic connection with the T2DREAM database. And entirely new data types (for example, metabolomic and proteomic data) arising from Opportunity Pool projects will be added to the T2DKP.

Ongoing work on tools and methods will result in the addition of many more interactive modules to the T2DKP. Researchers will be able to view PheWAS data; prune lists of variants by their linkage disequilibrium relationships; calculate credible sets and genetic risk scores with custom parameters; perform more versatile interactive burden tests; prioritize genes by pre-calculated association scores; overlay the positions of coding variants on protein structures to help assess their impact; and perform enrichment analysis on sets of loci to suggest pathways implicated in disease processes.

The Knowledge Portal platform developed for AMP T2D has already proved extensible to other complex diseases: in 2017, both the Cerebrovascular Disease and Cardiovascular Disease Knowledge Portals were launched. In the future, connections within the ecosystem formed by the T2D, Cerebrovascular, and Cardiovascular Portals will be improved, so that researchers can easily assess the impact of a variant or involvement of a gene for all of these related diseases. If funding and collaboration considerations allow, perhaps one day these Portals will merge into a single cardiometabolic disease genetics Knowledge Portal to accelerate the development of new therapeutics in this broader area.

Finally, the ultimate goal of this funding period is that by its end, the data generation, analysis, and interpretation will have facilitated the validation of multiple promising drug targets for further investigation. Given the rate of progress on multiple fronts, this seems a realistic goal. We hope that this unique collaborative environment will continue to accelerate T2D genetic research and will become a paradigm for other research communities.