AMP T2D Knowledge Portal Submitter's Guide to Sending Data to the DCC

Contents

Executive Summary

The AMP T2D Knowledge Portal is a web based portal in place for the type 2 diabetes scientific community that is transforming the way research communities share and visualize genetic data and facilitating new disease discoveries. In order to enable scientists to utilize these new tools on their data sets and increase the power of the data on the knowledge portal, the AMP T2D Data Coordinating Center (DCC) is bringing in new data sets for deposition into the AMP T2D Knowledge Portal. All data sets submitted to the DCC have the results from the analysis performed at the DCC uploaded to the knowledge portal that can be viewed by the knowledge portal users. Individual level data will not be shared on the knowledge portal. This document is a guide for studies that are interested in depositing their array, whole exome sequencing, or whole genome sequencing data into the portal. Other data types will be accepted in the future. Please see Figure 1 below for a brief overview of the submission process. Note that the association analysis will be done only on data sets with individual data.

Submitting your data to the DCC will be an interactive process between your analyst/PI and our analysis team. The data intake team at the DCC will be reviewing the QC and association analyses with the submitter before any data is uploaded onto the portal and working with the data submitter to resolve any issues found with the data. The analysis process is intended to be iterative and the data submitter and DCC will decide together the order and timeline for the association analysis.

Once an analysis is ready for submission to the knowledgebase, analysis will go live in the portal on the next release. Once the data is live on the portal, our publication policy comes into effect. The data will initially enter Early Access Period 1 for months 0-3 and Early Access Period 2 for months 3-6 months. During the first 6 months on the portal data will be flagged as Early Access and under the guidelines of the Fort Lauderdale principles. After the data has been on the portal for 6 months, the open access period will start for the data.

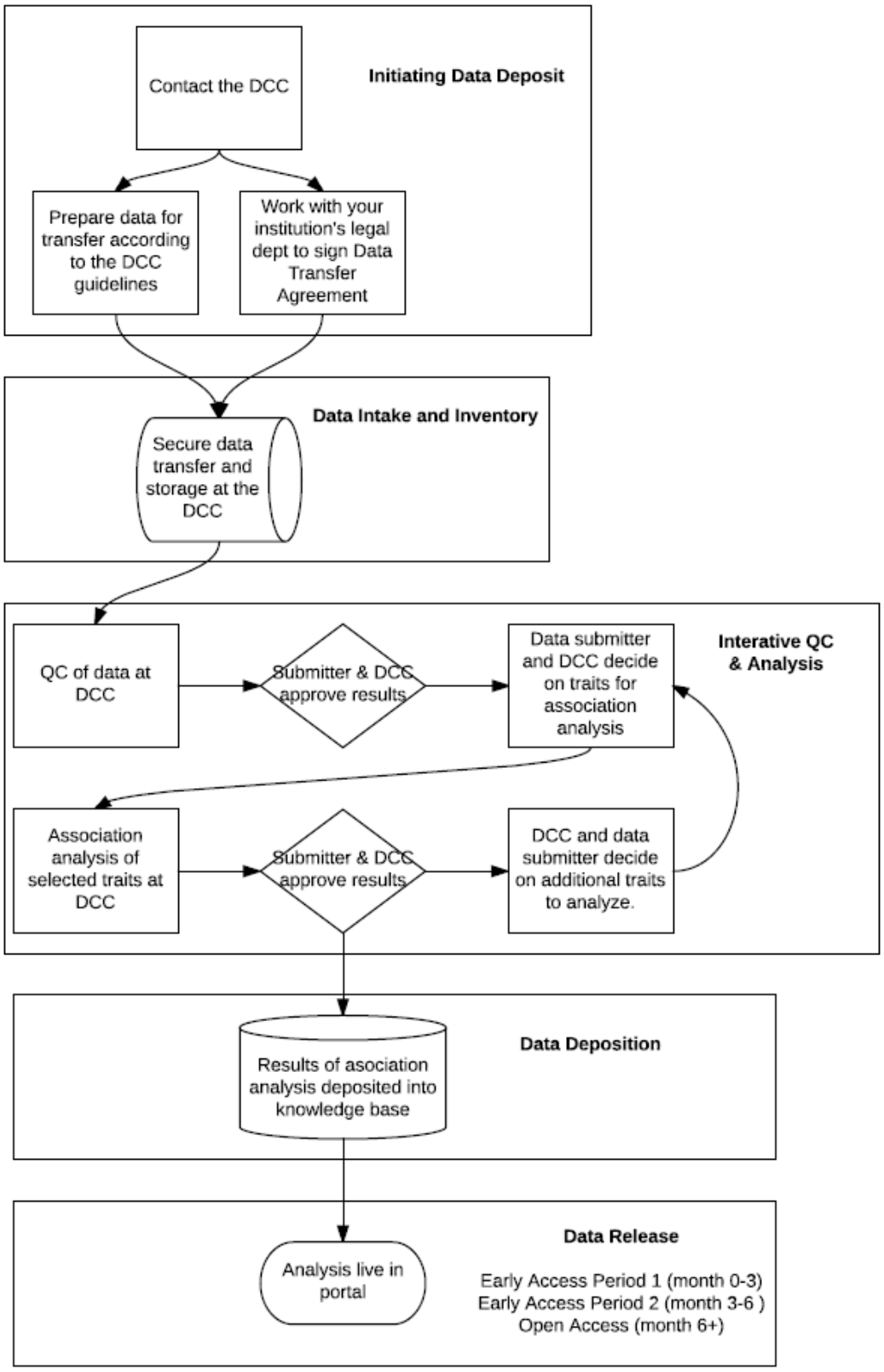

Figure 1. Overview of AMP T2D Knowledge Portal Data Submission at DCC

Summary of Milestones for Data Submission to the Data Coordinating Center

1. Signed DTA executed by both the submitter's institution and the Broad Institute, serving as DCC.
2. Submitter has prepared necessary files for transfer to the DCC.
3. Genetic data uploaded to secure transfer site. (Individual level data only)
4. Initial phenotype data uploaded to secure transfer site. (Individual level data only)
5. Pre computed analyses upload to secure transfer site. (Required for summary statistics submission and strongly recommended for individual level data submission)
6. Project info shared with submitter before being loaded to the portal.
7. Submitter and DCC approve project description.
8. Results of compliance check and analysis that will be shown on portal shared with submitter. (Summary statistics only)
9. Submitter and DCC approve deposition of summary level analyses on the portal. (Summary statistics only)
10. Analysis results shared with submitter. (Individual level data only)
11. Project information that will be loaded on the portal shared with submitter.
12. Submitter and DCC approve QC'ed data. (Individual level data only)
13. Submitter and DCC approve association analysis. (Individual level data only)
14. Submitter and DCC approve project information.
15. Analysis goes live on portal.

Contacting the DCC

To get started on this process, please reach out to the Data Coordinating Center at the Broad Institute by emailing us here: amp-dcc-data-submission@broadinstitute.com. Please tell us about the data set you'd like to submit and any concerns you have about depositing your data. A member of the data intake team will reply with additional information and guide you through the submission process.

Introduction

Welcome to the AMP T2D Submission Guideline! Bringing in new data to the knowledge portal increases the value of the AMP T2D Knowledge Portal for the type 2 diabetes research community and allows the submitter to see their data in the context of hundreds of thousands type 2 diabetes and control samples gathered from around the world. If you haven't yet, please check out the portal here: http://www.type2diabetesgenetics.org/. All you need to get started is a google log in.

This document outlines the process of submitting data to the AMP T2D Data Coordinating Center (DCC) at the Broad Institute in Cambridge, MA, USA and will serve as a guide to submitters throughout the process. The process mapped out below begins with getting your Data Transfer Agreement signed and ends with the deposition of your data in the portal. It reviews the process, roles, and responsibilities of the DCC and the data submitter. In addition to reviewing the information below, we encourage you to

reach out to your project manager with any issues or questions you encounter during your submission process. Each process has defined milestones that highlight significant progress in getting the data ready for deposition.

If you haven't started the process yet and are interested in depositing your data into the AMP T2D knowledge portal, please contact the DCC at the Broad Institute by emailing us here: amp-dcc-data-submission@broadinstitute.com. If you are unable to send your data to the USA for any reason, you have the option of submitting your data through a federated node. Please contact the DCC for more information.

AMP and the AMP T2D Knowledge Portal Overview

The Accelerating Medicines Partnership (AMP) effort is a public-private partnership between the National Institutes of Health (NIH), 10 pharmaceutical companies and multiple non-profit organizations that joined together to transform the way researchers identify and validate therapeutic targets for several diseases, including type 2 diabetes. To read more about the AMP initiative and to see who's involved, please visit: https://www.nih.gov/research-training/accelerating-medicines-partnership-amp/type-2-diabetes

The AMP type 2 diabetes (AMP T2D) consortium is a collaboration of a number of AMP funded investigators from around the world, including the Broad Institute, University of Oxford, and University of Michigan. The goal of the AMP T2D consortium is to create a knowledge portal using genetic and phenotypic data generated from type 2 diabetics and controls across multiple populations in order to bring forth discoveries in the genetic architecture of type 2 diabetes and to facilitate the development of new therapeutic targets for treating this disease. Using the genetic data collected from researchers around the world in an interactive web portal environment, researchers are able to ask questions from the data and see summary level results. You can also search for your gene, variant, or region of interest and see if any of the AMP T2D knowledge portal data sets have association for type 2 diabetes or related traits.

The AMP T2D knowledge portal will be continuing to work towards improving the value of the portal for the type 2 diabetes community. To this end, the AMP T2D consortium will be working to add new data sets to the portal and improve the web based tools used for analysis within the portal. We will be updating the community on our progress through the use of our mailing list and twitter feed by signing up on the home page of the portal: http://www.type2diabetesgenetics.org/home/portalHome.

Types of Data Requested for the AMP T2D Knowledge Portal

The Data Coordinating Center is currently able to accept array data, whole exome sequencing data, and whole genome sequencing data that is able to be transferred to the United States. We are building the capacity to accept other data types, such as gene expression, metabolomics, and epigenetic data.

Overview of Data Aggregation and Analysis Process

As a submitter to the knowledge portal, we know it's important for you to understand how your data will be handled once it is at the DCC. Only analytical results, and not individual level data, will be accessible through the portal. We anticipate that multiple versions of results, of increasing detail and harmonization with other datasets, will be released to the portal in time.

Each cohort/project being submitted to the knowledge portal will have the appropriate analytical results identified and harmonized with existing analyses in the portal through a collaborative process between an analyst at the DCC and an analyst at your institution. The analytical results that are prioritized will be dependent on the phenotype data available, the value the analysis adds to the knowledge portal, and any special requests made by the data submitter.

The analysis will be released to the portal in 3 stages: Early Access Phase 1, Early Acess Phase 2, and Open Access Period. Early Access Phase 1 gets the data analysis uploaded and available on the portal with limited QC. The subsequent revisions of the results will occur in Early Access Phase 2, which will aim to (a) address any inconsistencies identified by the initial harmonization process (b) apply more uniform QC across all datasets in the portal (c) compute additional statistics desired in the portal but not available in the initial version and (d) enable on-demand interactive analyses of your data. For these revisions we will require the original genotype and phenotype data. Additionally, we will also require data in as unprocessed a format as possible, in order to facilitate harmonization and quality control. Once the data is QC'ed and complete, the Open Access Period will begin for your data. We expect the timing between the start of Early Access Phase 1 to the beginning of the Open Access Period to last 6 months.

Policies and Data Use

We are committed to ensuring that collaborators submitting genetic data to the AMP T2D knowledge portal understand how the data will be used after transfer to the DCC at The Broad Institute. By sending your data to the Broad Institute for upload into the knowledge portal the data submitter and DCC are agreeing to the following:

1. Throughout this process, the Broad is committed to protecting your data, both in transit and while the data is in our servers.
2. We will only be able to receive de-identified level data that is able to be transferred and stored at the DCC. We will have options available for those who cannot submit data to the United States.
3. Individual data will be stored in our secure servers and only accessed for QC and analysis purposes related to the AMP T2D knowledge portal.
4. Individual data will never be posted directly to the portal. Only summary level metrics are available to portal users.
5. Summary level analysis of the submitted data will be posted to the knowledge portal and available to users. This includes p-values, odds ratio, minor allele frequency, effect, direction of

effect, allele frequencies across ethnicities, and other analyses that are deemed appropriate by the AMP T2D Knowledge Portal team and AMP T2D consortium.

6. Users of the portal will be able to create custom queries and view summary level results for those queries. This will include displaying results for specific projects/cohorts.
7. The Broad will QC and analyze your data for T2D and related traits in partnership with the submitter. This is a collaborative process so the submitter will get to view the analysis before it is uploaded to the portal.
8. The Broad may be sending genotype data submitted to the portal to the Michigan Imputation Server for imputation. This is a free service hosted by the University of Michigan and allows us to use the Haplotype Reference Consortium panel for imputation. The University of Michigan is a key member of the AMP T2D consortium that is funded by the NIH to develop the AMP T2D knowledge portal. For additional information on the imputation server, please visit: https://imputationserver.sph.umich.edu/index.html.

The policies related to the data in the AMP T2D knowledge portal, including data use for the knowledge portal users, can be found here: http://www.type2diabetesgenetics.org/informational/policies#.

Submitting Data that Cannot Enter the United States

Our AMP T2D funded collaborators at the University of Oxford are currently building a capability to ingest data, QC, and harmonize data at EBI. If data can't leave Europe or enter the United States you can still submit your data to the knowledge portal through this method. EBI will perform the same functions as the DCC and will work with you to

Data Transfer Agreement

Before we begin transferring data we need a signed and executed Data Transfer Approval (DTA). You will receive the DTA via email from the DCC project manager or you can find it on the knowledge portal here: http://www.type2diabetesgenetics.org/informational/policies. This document should be reviewed by your institution's legal counsel before signing and any edits made will need to be signed off by the legal counsel at the Broad. The document outlines that as a data contributor to the AMP T2D Portal, you agree to transfer your data to the DCC (Broad Institute) and you have the approval to do so. Although not covered in this document, a similar DTA will be necessary to transfer data to a Federated node in cases where the data cannot enter the United States.

Milestone:

1. Signed DTA executed by both the submitter's institution and the Broad Institute, serving as DCC.

Preparing for Data Submission to the Portal

While we work towards getting a DTA in place, the data submitter can begin the process of preparing their files for data submission to the DCC. The information below outlines the information we need to

get your data uploaded to the portal. For a summary table of the information needed, please see Table 1 below.

If your data is unable to leave your site or come to the Broad Institute, located in Cambridge, MA, USA, then depositing your data in a Federated Node will allow you to still contribute your data to the knowledge portal. Please contact the DCC for more information.

Required and Requested Files

Below are guidelines for the types and desired formats of datasets transferred to the DCC. As a general rule, we encourage you to submit as much data and as many results as possible, and to annotate your files with as much information as is feasible. This information will be extremely helpful as our analysts start the QC and analysis process on your data. Please note that we understand that different sites will have different data types and different abilities to transform among data formats, and we are thus happy to work with you to facilitate this process on a case-by-case basis.

1. **AMP DCC Data Intake Form.** This data is **required** in order to submit your data to the DCC. The form will be sent via email and please contact your project manager or amp-dcc-data-submission@broadinstitute.com if you have not received it. For additional details on the type of information needed, please refer to Appendix B.

2. **Analysis result files.** These files are **optional**, but any analytical results that you transfer will help us expedite and verify our analysis. Any number of files can be provided. For each file, the following is required

    - A tab- or comma- delimited file, with a header row followed by one row for every variant in the results file. The header row can have as many columns as possible. **Mandatory** columns include the chromosome, position, effect allele (with respect to which any phenotypic effect is measured), and non-effect allele. All alleles should be aligned to the forward strand of the genome, the version of which should be specified in the annotation data (see below). Additional desired columns include minor allele frequency, p-value of association with one or more traits, estimated odds ratio or effect size, case/control counts, and number of analyzed samples. If multiple statistics are available across multiple traits (e.g. T2D vs. glucose) or across multiple sample grouping (e.g. all samples vs. only samples of a given ancestry) they can be included in a single file or split across multiple files. The set of variants need not be identical across different result files.

    - Annotation data describing the meaning of each column are **required**. These should be human readable. The annotations can be embedded in the results file or provided as a separate document.

3. **Primary Genotype Data File Types.** In order to ensure the continued use of your data in the portal as demand for additional statistics and analyses grows, we request the following files encoding the genotypes of each sample. These genotype files will be used to compute statistics that are unavailable from the analysis files, which will be added to the portal in subsequent data versions.

    - Genotype files in VCF or PLINK format are **required**. We will accept either format, provided that strand information is clearly annotated. Note that VCF files have a clear distinction between reference and alternate alleles, while alleles can be flipped by some plink analyses.

        a. The VCF file format is available at XXX.

        b. Information about the PLINK file format is available at XXX. We recommend transferring bed/bim/fam files, which can be created by PLINK.

    - Lists of QC+ samples and variants that were advanced to your final analysis are **optional**. Providing these will ensure that we can recompute statistics concordant with those that you produced in your analysis. If you do not provide them, we will perform our own QC which will likely be similar (but not identical) to yours.

    - Documentation of your original analysis plan is also **optional**. Any human readable document describing the motivations of your analysis, the statistical methods employed, and any parameter settings will also help us to replicate your analysis. A methods section of a paper, if sufficiently detailed, will also suffice.

4. **Intensity files for SNP array data.** Ultimately, it may be necessary for us to have access to the raw data used to call genotypes. This will assist with quality control (for example, examining evidence that a rare variant has accurate genotypes), as well as harmonization (for example, ensuring that all variants are called using similar procedures). Thus, although not essential for the first version of your data to appear on the portal, the following files are **required** to complete the data transfer process in its entirety.

    - Raw intensity files (idat or the equivalent). For SNP array genotyping data, any file format that lists normalized X/Y intensity values for each sample is acceptable. When submitting IDAT, please remember to send both of the intensity files for each sample.

        a. Example file formats accepted by the Sanger for a similar project are at XXX

        b. A guide for the file formats used by zCall (a clustering algorithm for exome chip) is available at XXX.

    - Cluster and manifest files to accompany the raw intensity files. For Illumina IDAT files, these two files are required for the necessary downstream analysis. They should be available from and familiar to the platform that produced your original genotype calls. The manifest file

describes the samples that were genotyped; the cluster file records any information that was used to better cluster the intensities for each SNP.

5. **Read files for sequencing data.** Similar to intensity files for SNP array data, read files are **required** for sequencing data. We will use these to run "joint variant calling" across all samples at the DCC, for maximum sensitivity and accuracy of variant calls. Since variant call sets from sequence data include novel variants and alleles, re-processing raw data is even more important in sequencing experiments than SNP array experiments; the ExAC paper (available at XXX) outlines some of the rationale for this.

    - BAM or CRAM files for each sample are required for sequencing experiments. These files are the standard format for storing read data and should be produced by your sequencing platform. We would prefer raw, unaligned BAM files.

        a. Information on the BAM file format is available at XXX

        b. BAM files can be created from FASTQ files, as described at XXX.

6. **Phenotype Data.** This is **required** alongside submission of genotype and/or sequencing data. The official document with full instructions will be emailed to you. For an idea of the variables requested, please see Appendix C. If you have a specific variable not in this list, but relative to type 2 diabetes or related conditions let us know and we can include it for your submission.

For a summary view of what is needed for your data submission, please see table 1 below.

Table 1. Summary of files accepted for data submission into the AMP T2D Portal

File Type	Genotyping Submission	Sequencing Submission
AMP DCC Data Intake Form	Required	Required
Analysis Results	Optional	Optional
Annotation Data	Optional	Optional
Genotype Files (VCF or PLINK)	Required	Required (VCF)
List of QC+ samples and variants	Optional	Optional
Analysis Plan Documentation	Optional	Optional
Raw Intensity Files	Required	N/A
Cluster File	Required	N/A
Manifest File	Required	N/A
Sequencing Read Files (BAM or CRAM)	N/A	Required
Phenotype Data	Required	Required

Milestone:

1. Submitter has prepared necessary files for transfer to the DCC.

Overview of the Data Intake, Analysis, and Deposition Process

When you are ready to start submitting files to the DCC for deposition into the AMP T2D portal, email amp_dcc_data_submission@broadinstitute.org and we will set up a secure transfer portal for you to upload your files. We will be using an ASPERA site, which will come with detailed instructions on how to upload the files. Once the ASPERA site is created, we have 30 days before the site expires to upload data. If it becomes necessary to extend that timeline please let us know so we can extend the life of the ASPERA site.

The data transfer process it outlined by step below. For a full picture of data intake to deposition, please see Appendix A.

Data Transfer

The data transfer process starts once the submitter and DCC at the Broad Institute have all necessary documentation in place and are ready to begin physically transferring the data to the DCC. During these steps, if individual level data is being provided, the data submitter will transfer the phenotypic and genetic data to the portal. This includes the raw data and any available pre computed analyses. For sites where we are receiving summary statistics only we ask for the pre computed analyses to be sent to the DCC. Please see Figure 2 below for an overview of the data intake process at the DCC.

Regardless of which type of submission being sent, we ask that each site completes a data intake form, as noted above. The purpose of the form is to inform the DCC of the data being submitted and to help us create a project/cohort description for this data on the portal.

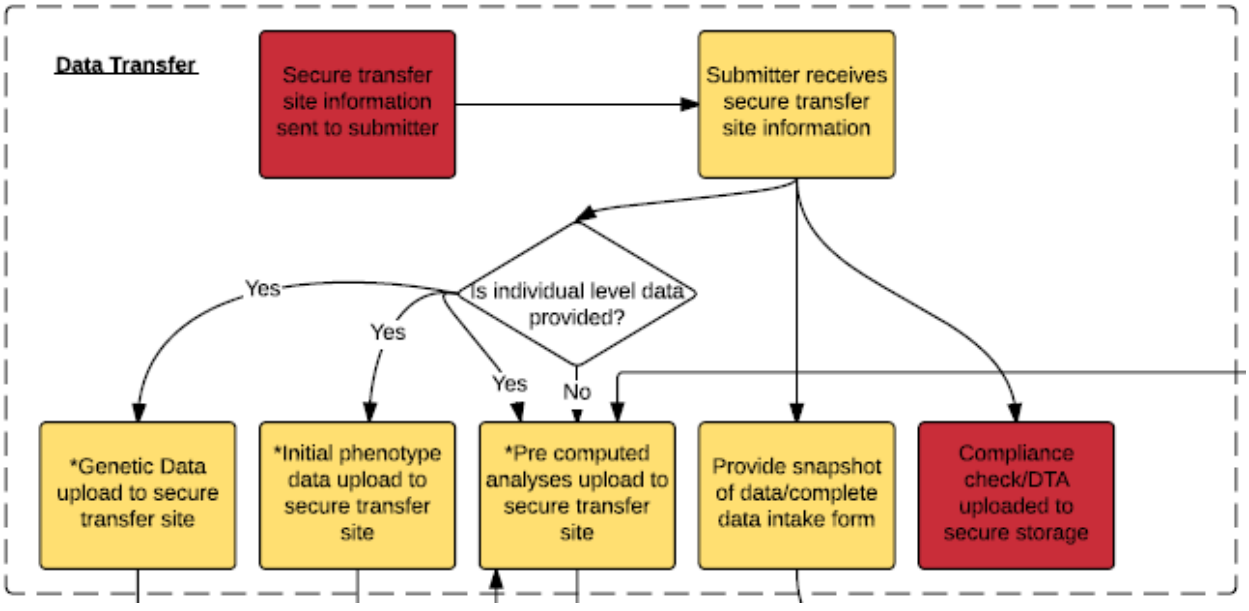

Figure 2. Data Transfer Process at DCC

Milestones:

1. Genetic data uploaded to secure transfer site. (Individual level data only)
2. Initial phenotype data uploaded to secure transfer site. (Individual level data only)
3. Pre computed analyses upload to secure transfer site. (Required for summary statistics submission and strongly recommended for individual level data submission)

Description of Project/Cohort

Each project and cohort with data included in the AMP T2D portal will have a description of the project and/or cohort that is submitting data. This description will be created by the Content Manager at the DCC using the project information provided by the submitter on the Data Intake Form. During this process, the submitter will have the opportunity to provide feedback on the description of their study.

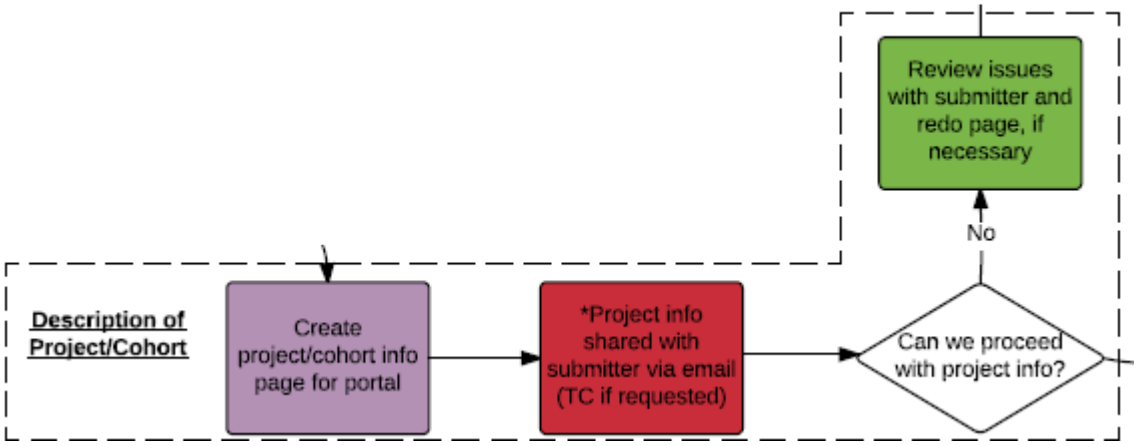

Figure 3. Description of Project and Cohort Information Submission Process at DCC

Milestones:

1. Project info shared with submitter before being loaded to the portal.
2. Submitter and DCC approve project description.

Summary Statistics Only

If you are not able to share raw data with the portal for some reason, the portal can accept summary level statistics that can be posted to the portal. In this instance, the DCC would take the summary level information that you have generated then securely store the data and perform a data compliance check. Once the compliance check has been completed, the DCC will share results with the submitter and confirm that we can proceed with depositing the data to the portal.

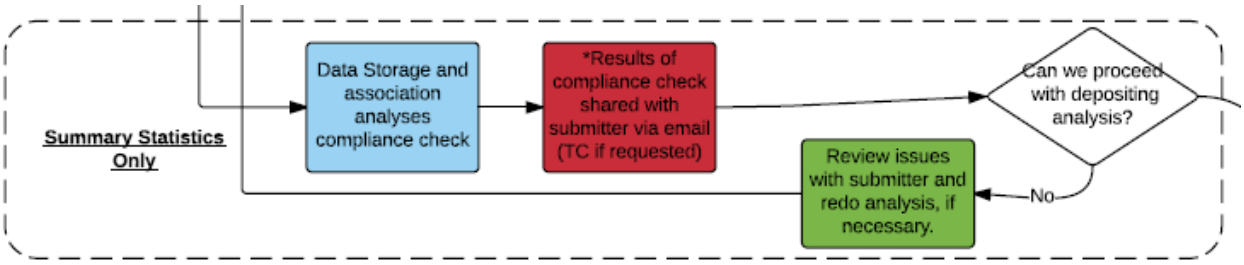

Figure 4. Summary Statistics Only Data Submission Process at DCC

Milestones:

1. Results of compliance check and analysis that will be shown on portal shared with submitter.
2. Submitter and DCC approve deposition of summary level analyses on the portal.

Data QC and Analysis at DCC

Data sets with individual level data being submitted to the DCC will undergo secure data storage, QC, and association analysis at the DCC. During this process the DCC will work with the data submitter to create an analysis plan that will be used to drive the future analyses and create data sets within the project/cohort that will be deposited into the portal. A data set in this context refers to a specific set of samples paired with specific phenotype(s). We expect each data submission to contain a number of data sets and we will work with the data submitters to prioritize the data sets for submission to the AMP T2D portal. Once the analysis is completed the DCC will reach out to the submitter and review the results of the analysis. Results will not be uploaded to the portal until both the data submitter and DCC affirms that the data is ready to share.

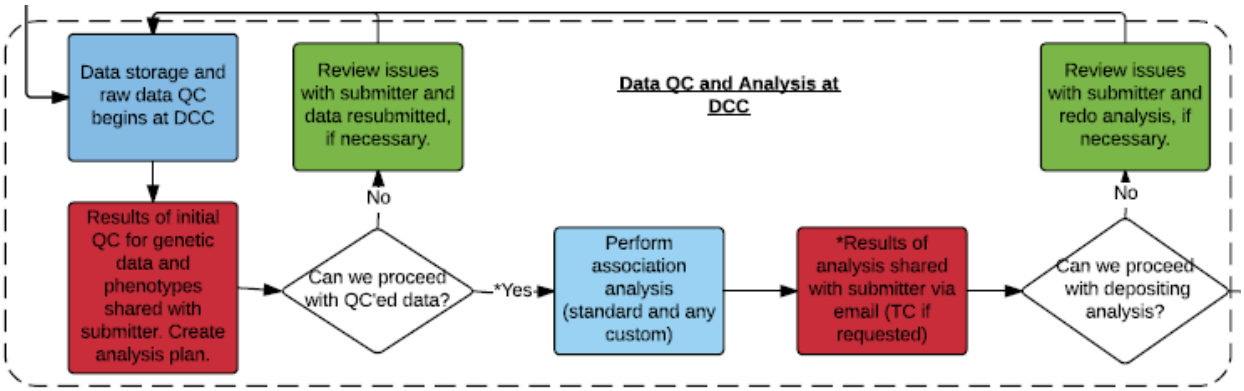

Figure 5. Data QC and Analysis Process for incoming data to the DCC

The DCC has compiled a list of standard single variant association analyses that will be used as a guide for creating the analysis plan with the submitter. Each analysis plan will be unique to each site, depending on the phenotype variables that are available and the value each analysis adds to the portal.

Milestones:

1. Analysis results shared with submitter.
2. Project information that will be loaded on the portal shared with submitter.
3. Submitter and DCC approve QC'ed data.
4. Submitter and DCC approve association analysis.
5. Submitter and DCC approve project information.

QC Process at the DCC

The QC process at the DCC is vital to harmonizing the data being added to the AMP T2D Knowledge Portal. The goal of our QC is to identify artifacts, ensure statistics can be computed consistently, and helps the DCC understand the data being submitted. This process is undertaken on individual level data that is provided to the DCC by an analyst who is working with all data being loaded into the knowledge portal and performs the QC using an automated and consistent process.

The analyst at the DCC will be computing metrics adjusted for ancestry and other confounders and then exclude outlier samples, which are potentially artifacts. The QC completed for data destined for the knowledge portal tends to be conservatives, since we are aiming to ensure high quality data for users. Once this QC has completed, we will provide a report to share with the data submitters. For full details on this process please see Appendix D.

Association Analysis Process at the DCC

Association analysis at the DCC is an interactive process between the analyst at the DCC and the analyst at the submitting site. The initial analysis performed will consist of a set number of traits decided upon by the DCC and the data submitter. As a guide for our submitters, the DCC recommends focusing on some initial traits of relevance to type 2 diabetes for the initial analysis done at the DCC. Please see Table 2 for a list of recommended traits.

Table 2. Standard T2D traits for possible DCC association analysis

Categories of traits	Example related phenotype variables
Type 2 Diabetes status	T2D status, T2D age of diagnosis
Cardio metabolic	Systolic blood pressure, Diastolic blood pressure, Hypertension status
Lipids	HDL cholesterol, LDL cholesterol, Triglycerides, Total cholesterol
Glycemic	Insulin, glucose (2hr, fasting, and/or random) HbA1C,
Anthropomorphic	BMI, age, weight, waist hip ratio
Kidney Function	Creatinine, Urinary albumin

Data Deposition and Release

Once the analyses and the project/cohort description have been reviewed and approved by both the data submitter and DCC, the data will be deposited onto the knowledge portal base before going live on the portal.

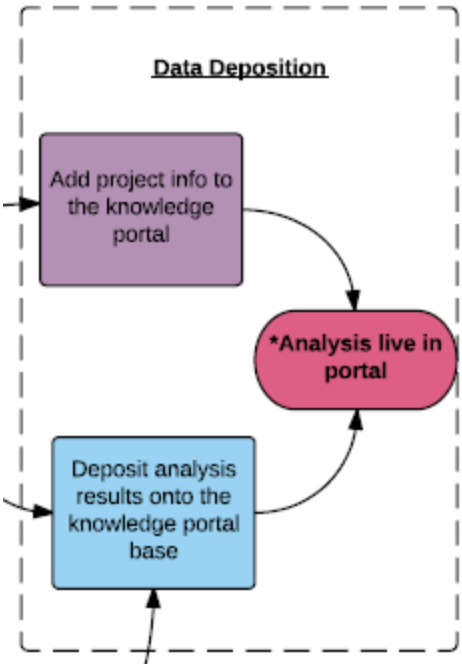

Figure 6. Data Deposition Process at DCC

The data will go live with the next quarterly portal release, occurring in February, May, August, and November.

Milestone:

1. Analysis goes live on portal.

Publication Policy

Once your data is live on the knowledge portal, submitters are protected by a 6 month early access period that is subject to the guidelines of the Ft. Lauderdale principles. This 6 month period is broken down into a 3 month Early Access Phase 1, where data is live on the portal with limited QC and a 3 month Early Access Phase 2, where the data is fully integrated into the portal. All data in either of the Early Access Periods will be flagged to knowledge portal users who are viewing the data. Please see Figure 7 for the scheduled data releases.

For more information on the Ft. Lauderdale policies, please visit: https://www.genome.gov/pages/research/wellcomereport0303.pdf.

Submitted data will be made live on the knowledge portal over a number of data set freezes. Since we will be running association analysis over time and adding to the portal, each data set will be defined as a set of genetic data associated with specific traits. Any analysis additional traits, samples, or data will be

considered a new data set and will start again in the Early Access Period 1. For example, if for the initial analysis the data submitter and DCC chose to run an association analysis on 3,000 Exome chip samples using type 2 diabetes status that would equal to one data set and would start the Early Access Period on the next scheduled release of the portal. If the same 3,000 exome chip samples were then analyzed later for BMI, fasting glucose, and fasting insulin that would create a new data set that would start in the Early Access Period, even if the same samples analyzed for type 2 diabetes have the analysis in the open access period.

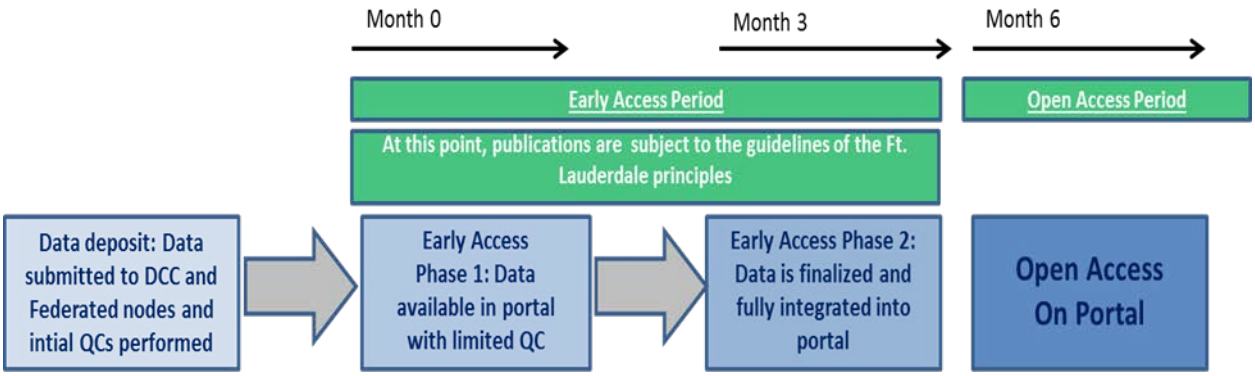

Figure 7. Scheduled AMP T2D Portal Releases for Data Submission

Appendix A: AMP DCC Data Intake to Data Deposit in AMP T2D Knowledge Portal

Figure 8 is a complete flowchart outlining the DCC's data intake process, starting at the point where the data is legally and physically prepared to be transferred by the submitter to the DCC and ending with the data being live in the portal. This document contains 5 subsections of work that is grouped together to create the larger process. The subsections are discussed in more detail in the main document.

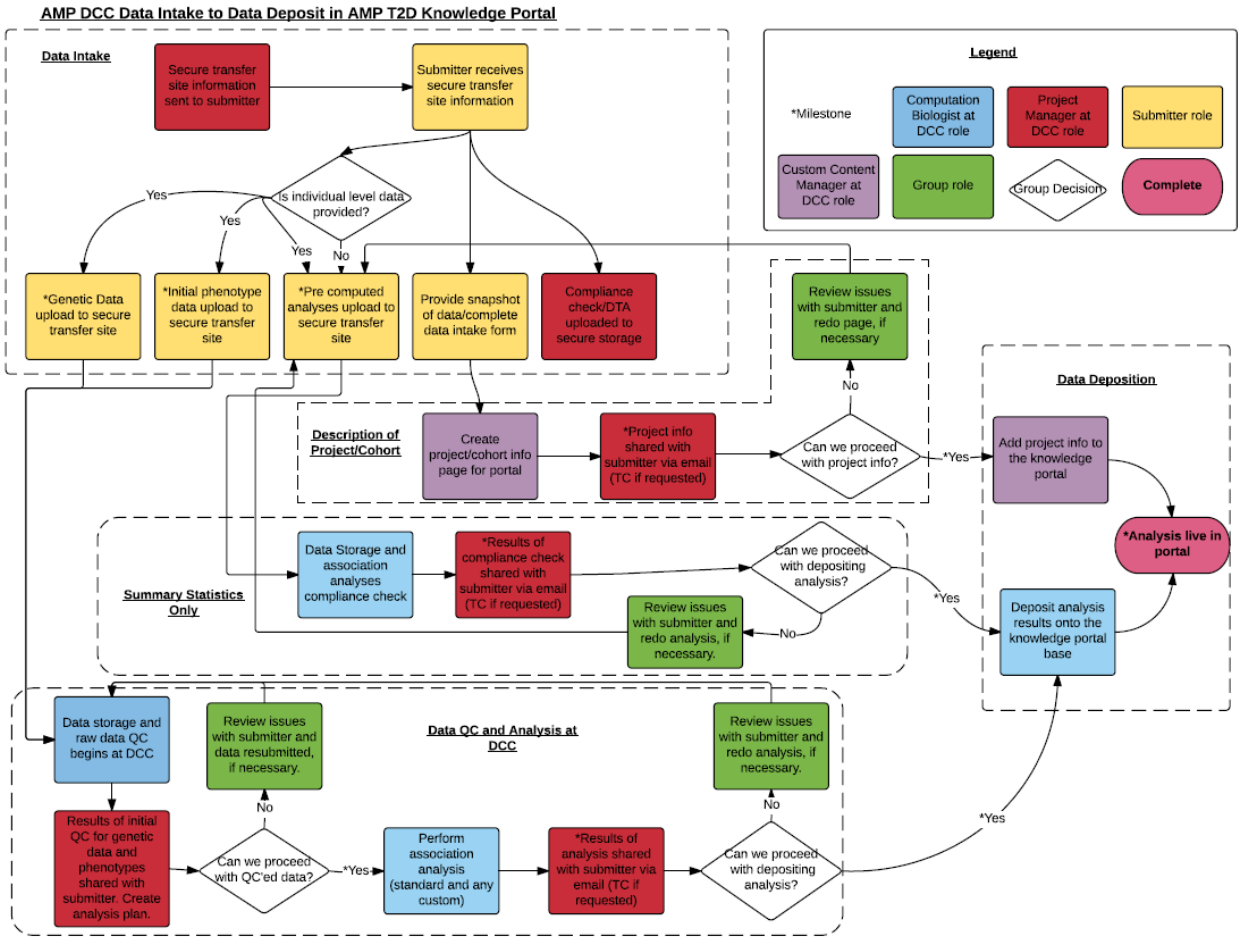

Figure 8. Flowchart of AMP DCC Data Intake to Data Deposition in AMP T2D Knowledge Portal

Appendix B: AMP DCC Data Intake Form

For a copy of the AMP DCC Data Intake Form to complete please contact your project manager or amp-dcc-data-submission@broadinstitute.com. In order to give you an idea of the type of information needed, we have included a screen shot of the information below. The information will be given for the project on the first tab (see Figure 9) and by cohort(s) in the following tabs (see Figure 10).

Data Submission Information	*Please list all data types, technologies, samples and file types that are being uploaded as part of this submission.			
Data submission type (Genotyping or Sequencing)	**Data technology**	**# of samples**	**File type(s) being submitted**	**Check sums**
Site Information	**Submitter Information**	**Instructions**		
PI(s)				
PI email(s)				
Submitter Name(s)				
Submitter Email(s)				
Total # of files being submitted				
Project Name				
Project Description				
Funding acknowledgements				
Project logo (image file preferred)		*Please attach image file in reply email or send the link to the website where we can find the logo		
Project website link				
Timing of submission				
Ethical considerations				
Special analyses request		*List any special requests for analyses to be performed at the DCC.		
dbGAP ID		*If the data is available in dbGAP, please give us the phs ID		
EGA IDs		*If the data is available in EGA, please give us the project ID		

Figure 9. AMP DCC Data Intake Form Project Information

Site Information	**Submitter Information**	**Instructions**
Cohort Name		
Cohort Description		
Cohort logo (image file preferred)		*Please attach image file in reply email or send the link to the website where we can find the logo
Cohort website link		
Sample Collection Site		
Related publications (PM ID preferred)		*If PM ID not available, please include the Authors, Title of Article or Abstract, Journal Name, Year;Volume:Start page-End page.
Cohort specific publications (PM ID preferred)		*If PM ID not available, please include the Authors, Title of Article or Abstract, Journal Name, Year;Volume:Start page-End page.
# of T2D cases		
# of T2D controls		
Ethnicity		*List all ethnicities being submitted
Case and control selection criteria		
dbGAP ID		*If the data is available in dbGAP, please give us the phs ID
EGA IDs		*If the data is available in EGA, please give us the project ID

Figure 10. AMP DCC Data Intake Form Cohort Information

Appendix C: Phenotype Submission

The AMP T2D consortium has determined a number of traits that will be useful for understanding your data and performing relevant analysis that can be shared on the knowledge portal. We ask that you submit any of these variables that are available for your data and also please let us know if you have a unique variable that we should be including. This list is meant for information purposes only. Please see the AMP Phenotype Variable Info sheet emailed to you for additional instructions and information.

Table 3. AMP T2D Knowledge Portal Phenotype Variables

Category	Variable	Format
ID variables	Study ID	character
ID variables	Sample ID used in genotype data set (if different)	character
ID variables	dbGaP sample ID (if existing)	character
ID variables	Study ID of father	character
ID variables	Study ID of mother	character
Demographics	Race	character
Demographics	Race - open text description	character
Demographics	Ethnicity	character
Demographics	Sex	Please code values as "Male" and "Female"
Demographics	Year of birth	4-digit integer
Type 2 Diabetes (T2D) status variables	T2D status based on self-report (1=T2D; 0=not T2D)	integer (1=T2D; 0=not T2D)
Type 2 Diabetes (T2D) status variables	T2D status based on history of healthcare provider diagnosis (1=T2D; 0=not T2D)	integer (1=T2D; 0=not T2D)
Type 2 Diabetes (T2D) status variables	T2D medication status (1=Yes; 0=No)	integer (1=yes; 0=no)
Type 2 Diabetes (T2D) status variables	T2D status based on fasting glucose level (1=T2D; 0=not T2D)	integer (1=T2D; 0=not T2D)
Type 2 Diabetes (T2D) status variables	T2D status based on HbA1c (1=T2D; 0=not T2D)	integer (1=T2D; 0=not T2D)
Type 2 Diabetes (T2D) status variables	Glucose tolerance status based on oral glucose tolerance test (OGTT)	character
Type 2 Diabetes (T2D) status variables	T2D status defined in a way other than one of the approaches above (1=T2D; 0=not T2D) - e.g. a combination of the above that can't be separated into individual variables	integer (1=T2D; 0=not T2D)
Type 2 Diabetes (T2D) status variables	T2D status with unknown definition (1=T2D; 0=not T2D) - e.g. where a T2D status variable is available but there is not documentation on how it was defined	integer (1=T2D; 0=not T2D)
Type 2 Diabetes (T2D) status variables	T2D treatment with insulin or analogs	integer (1=yes; 0=no)
Type 2 Diabetes (T2D) status variables	T2D treatment with non-insulin medication	integer (1=yes; 0=no)

Type 2 Diabetes (T2D) status variables	T2D age of diagnosis (for those that are affected) (years)	nn.nnn
Type 2 Diabetes (T2D) status variables	Time interval, in years, between diagnosis of diabetes and beginning of treatment with insulin	integer
Type 2 Diabetes (T2D) status variables	This is an open text variable to indicate types of diabetes other than Type 2 (or unclear if Type 2). Examples include: "Type 1 diabetes", "MODY", "LADA", "Gestational diabetes", "Diabetes known to be caused by other processes such as cystic fibrosis, hemochromatosis or pancreatic surgery", "Diabetes status only available during pregnancy".	text
Blood biomarkers	Fasting plasma glucose (mmol/l)	nn.nn
Blood biomarkers	Fasting insulin (mU/l)	nnn.n
Blood biomarkers	OGTT 2-hour fasting glucose (mmol/l)	nn.nn
Blood biomarkers	OGTT 2-hour fasting Insulin (mU/l)	nnn.n
Blood biomarkers	Random glucose (i.e. not fasting or unknown fasting) (mmol/l)	nn.nn
Blood biomarkers	Fasting C-peptide (nmol/l)	nn.nn
Blood biomarkers	HbA1c (fraction, %)	nnn.n
Blood biomarkers	HbA1c (mmol/mol)	nn.nn
Blood biomarkers	Glutamic Acid Decarboxylase Autoantibodies (GAD Ab)	integer (1=positive; 0=negative)
Blood biomarkers	Islet Cell Autoantibodies	integer (1=positive; 0=negative)
Blood biomarkers	Anti-insulin Autoantibodies	integer (1=positive; 0=negative)
Blood biomarkers	ZNT8 Autoantibodies	integer (1=positive; 0=negative)
Blood biomarkers	Serum creatinine (umol/L)	nnn.n
Blood biomarkers	Adiponectin (ug/ml)	nn.nn
Blood biomarkers	Leptin (ng/ml)	nnn.n
Blood biomarkers	Total cholesterol (mmol/l)	nn.nn
Blood biomarkers	LDL cholesterol (mmol/l) (if measured directly, missing if not)	nn.nn
Blood biomarkers	Calculated LDL cholesterol (mmol/l) (using Friedewald equation)	nn.nn
Blood biomarkers	HDL cholesterol (mmol/l)	nn.nn
Blood biomarkers	Triglycerides (mmol/l)	nn.nn
Blood biomarkers	Any lipid lowering medication status (1=yes, 0=no)	integer (1=yes; 0=no)
Blood biomarkers	Statin medication status (1=yes, 0=no)	integer (1=yes; 0=no)
Anthropometry	Height (centimeters)	nnn.n
Anthropometry	Weight (kg)	nnn.n
Anthropometry	Hip circumference (centimeters)	nnn.n
Anthropometry	Waist circumference (centimeters)	nnn.n
Blood pressure and hypertension	Systolic blood pressure (mmHg)	nnn.n
Blood pressure and hypertension	Diastolic blood pressure (mmHg)	nnn.n

Blood pressure and hypertension	Hypertension status (1=yes, 0=no)	integer (1=yes; 0=no)
Blood pressure and hypertension	Hypertension medication status (1=yes, 0=no)	integer (1=yes; 0=no)
Urine measures	Urinary creatinine (mg/dL)	nn.nn
Urine measures	Urinary albumin (mg/dL)	nn.nn
Urine measures	Urinary albumin to creatinine ratio (mg/g)	nn.nn
Smoking status	Current smoking status (1=yes, 0=no)	integer (1=yes; 0=no)
Smoking status	Ever smoking status (1=yes, 0=no)	integer (1=yes; 0=no)
Reproductive and exogenous hormone use	Menopausal status	character
Reproductive and exogenous hormone use	Current use of *any* female hormones (1=yes, 0=no)	integer (1=yes; 0=no)
Reproductive and exogenous hormone use	Current use of, specifically, peri- or post-menopausal hormone use (i.e. not including contraceptives) (1=yes, 0=no)	integer (1=yes; 0=no)

Appendix D: Detailed Overview of QC Process at the DCC

Quality Control Process at the DCC

All data submitted to the DCC will be processed through comprehensive sample and variant quality control algorithms to promote harmonization with existing data on the portal. Since genotype data is likely to exhibit unique patterns of ancestry and classes of variants, we have developed algorithms for detecting major lines of ancestry and for identifying outliers among various sample metrics. Sample QC will be performed using bi-allelic variants only.

Initial Data Review

Initially, when the DCC receives your data, it will be checked for duplicates and any cryptic relatedness that may result from contamination or data collection errors. Duplicates and cryptic relatedness will be identified using a combination of pairwise identity by descent and a robust algorithm for calculating pairwise kinship in the presence of population stratification. Should any concerns arise, the submitter may be contacted in order to investigate possible causes and issues that might be corrected prior to continuing with sample QC.

Ancestry Inference, Clustering, and Outlier detection

After an agreement to proceed with QC, we will infer major lines of ancestry. Our approach consists of projecting your data onto principal components derived from a collection of common ancestry informative variants in 1000 Genomes Project data. The PCs are then used as features in a Gaussian Mixture Modeling algorithm to cluster them according to their ancestry. Any samples that cannot be included in any of the subsets due to their unique ancestry or bad genotyping, are flagged as outliers.

Sample Metric Outlier Detection

During clustering, metrics for each sample will be calculated. Which metrics are calculated will vary depending on the type of data received. Some of the more recognizable metrics are transition/transversion rate, call rate, and the number of singletons called. For each sample metric, we will calculate the residuals resulting from regressing the metric on principal components of ancestry. Then we will calculate principal components on those adjusted metrics. Gaussian Mixture Modeling is employed again at this stage, both on the principal components of the adjusted metrics, and on each of the individual adjusted metrics. Any samples that do not cluster using these two approaches will be flagged as outliers.

Pedigree Reconstruction

If your data is found to have pairs of related samples, pedigree reconstruction will be performed.

QC Report

Upon completion of sample QC, a report will be provided to the submitter to facilitate the creation of a suitable analysis plan.

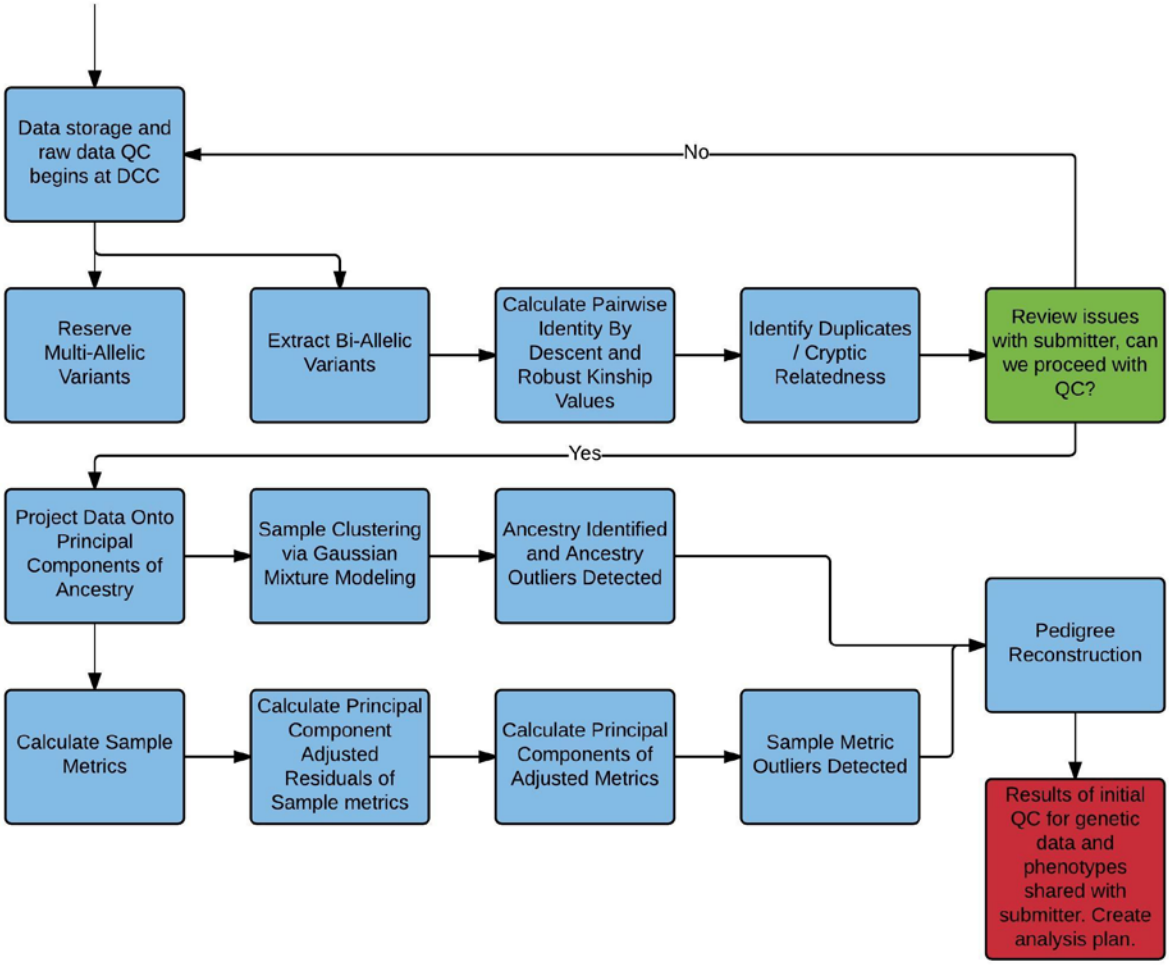

Figure 11. AMP T2D Quality Control Process