IX. ANALYSIS AND DATA DISTRIBUTION

A. INTRODUCTION

The goal of the Type 1 Diabetes Genetics Consortium (T1DGC) is to organize international efforts to identify genes that determine an individual's risk of type 1 diabetes. Statistical genetic analyses will be performed on appropriately powered data sets to identify regions linked to type 1 diabetes and determine how these linked regions act and potentially interact. In addition, the Consortium will perform a genome-wide association scan (GWAS) in cases and controls to identify regions/genes that are associated with risk for type 1 diabetes. This document describes plans and policies related to these statistical analyses.

B. DATA INVENTORY

The Type 1 Diabetes Genetics Consortium will accumulate the following data holdings:

- phenotype data from newly recruited cohorts
- phenotype data from cohorts established prior to the T1DGC
- results from T1DGC analyses (of DNA and serum) that have been directed by its Steering Committee (for existing collections as well as the current collection)
- results from T1DGC study-directed genotyping activities
- results from analyses of T1DGC specimens that have been conducted by individual investigators through access granted by the T1DCG Access Committee
- process data (*e.g.*, records of shipments and aliquots) concerning the collection and storage of specimens

In order to facilitate use of study data, the T1DGC Coordinating Center will maintain a web-based reporting system that will describe data in the T1DGC archives. Reports of the current holdings of the T1DGC data repositories are maintained on the study web site (http://www.t1dgc.org). Included among these reports is a summary of data sets and samples available to investigators. Data and samples are available to Contributing Investigators, Consortium Members and Non-Consortium Members according to the timeline outlined in the *T1DGC Policy Governing Access to Study Repository Samples and Data*.

C. STANDARD DATA MANAGEMENT AND PERFORMANCE REPORTS

The T1DGC web site will be used to display dynamic and static data reports used for managing the study and evaluating its performance. Reports will be divided into four nodes: recruitment, quality control, statistical and access.

1. Recruitment Reports

The recruitment reports will track the progress of the overall study and individual clinics in meeting recruitment goals. Cumulative totals versus goals over time will be displayed.

2. Quality Control Reports

Quality control reports will track the completeness of data entry, times between data collection and data entry, and response times to editing queries. Laboratory quality control reports will include information regarding consistency between quality control samples, results from laboratory tests and the frequency of any lost specimens.

3. Statistical Reports

Statistical reports will describe the information received from study-directed genotyping projects.

4. Access Reports

Access reports will describe the availability of T1DGC data and samples. Investigators can use these reports when requesting resources from the Consortium.

D. OVERALL ANALYSIS PLAN

1. Genetic Data

A genome-wide linkage scan and a genome-wide association scan will be performed to identify chromosomal regions and genes that influence risk of type 1 diabetes. The T1DCG Coordinating Center will perform linkage analysis using nonparametric methods to detect linkage between the panel of genetic markers and disease. The basic approach to linkage analysis has been described extensively, and is implemented in GeneHunter (1) and other software variants. The strength of the approach is that it is relatively model-free, requiring no prior knowledge of segregation parameters. The method requires the probability that each pair of relatives in a pedigree is identical by descent (IBD) at a given marker be known. Association analyses will be used to determine positional candidate genes. First, in the affected sib-pair (ASP) families, family-based association analysis, such as the transmission disequilibrium test (TDT) (2) and several TDT-like tests will be applied to the "fine mapping" markers to further localize genes within regions detected in the initial genome scan. We anticipate that the adaptation of the TDT to use the unaffected sibling (3) will be more suitable for our analyses because it will allow us to maintain the tests of linkage as well as linkage disequilibrium. Second, trio families containing African and Asian families using trans-ethnic mapping will be used to further explore associations detected in the affected sib-pair families. Standard association analyses using cases and controls are an attractive complement to family-based approaches to genetic association incorporating large scale (~500,000) SNP genotyping. Additional linkage disequilibrium mapping methods will be utilized to detect association between disease status and polymorphisms in the candidate genes using case-control samples available to the T1DGC.

2. Linkage

Upon completion of the genome scan and evaluation of marker distribution (PedCheck) and pedigree structure (PREST), a series of additional analyses will be performed prior to formal linkage. The nonparametric affected relative pair linkage analyses implemented in the GeneHunter package and its derivatives are an extension of the strategy developed by Kruglyak and Lander (4). Multipoint analysis is conducted by using information from adjacent loci to obtain more precise estimates of allele-sharing at a given chromosomal position. Accurate estimates of allele frequencies are required for linkage analysis when parents are missing or not genotyped. Because it is anticipated that almost all of the parents will be available for genotyping, we expect that differences in marker allele frequency will have minimum effect.

Many loci showing "suggestive" evidence for linkage (and even some with "significant" evidence for linkage) may be false positives. One method that should help to differentiate among "true" and "false" signals of linkage will be to perform a multilocus, or oligogenic, analysis in which we include all loci for which evidence for linkage was initially obtained. Such an analysis is performed as a straightforward extension of the nonparametric linkage (NPL) regression approach by including in the model locus-specific effects at multiple loci. Joint analysis of these suggestive loci will have two main benefits. First, as noted above, it should help to eliminate false positives since joint estimation should yield more accurate estimates of effect size. Second, it should increase the power to detect linkage since conditional testing using profile lod scores should maximize the relative signal-to-noise ratio of true linkages. This approach will results in both the elimination of false positives and the increased signal of true linkages can occur. For the linkage analyses, all sib-pairs will be combined. Specific genetic risk factor markers (*i.e.*, HLA, *INS, and CTLA4*) will be used as modifying effects in the models, in order to be determining if specific loci play a more significant role in certain genetic subgroups.

3. Family-based Association

Family-based association analyses will be performed using "fine mapping" SNPs to further localize susceptibility loci detected in the genome scan, candidate regions, or candidate genes. The TDT, as originally designed, is a test of both linkage and association that evaluates if affected offspring inherits a particular allele more often than would be expected by chance (5-6). It eliminates the effect of population substructure because both cases and controls are obtained from the same parent. The TDT has been extended from biallelic to multiple allele markers (7), from parent-child trios to discordant sibships (8), and from qualitative traits to quantitative traits (9). Although the most popular method is TDT as proposed by Spielman and Ewens (6), work by Schaid (10) has compared a series of statistics for analysis of association and suggests alternative methods for use in absence of clear transmission models. These statistics include a generalized TDT (GTDT), a series of scoring statistics (including GEN, a generalized score when the action of alleles is unknown), and "standardized" scoring statistics. As shown by Schaid (10), when the number of high-risk alleles is unknown and the effect of the marker alleles (dominant versus recessive) has yet to be resolved, the GTDT statistic appeared most powerful. In addition to the standard TDT analyses, haplotype analyses will be performed.

4. Case-Control Association

Using a large population of cases and controls with large SNP genotyping panels, genome-wide evidence of association can be determined using a multi-stage experiment. The first stage of this experiment will be to perform an association scan using ~500,000 SNPs in 4,000 cases from the JDRF/WT British case collection and 2,500 controls from the British 1958 Birth Cohort (B1958BC). This approach will augment the existing Wellcome Trust Case-Control Consortium (WTCCC) study of type 1 diabetes - a design consisting of 2,000 cases (from the JDRF/WT British case collection of over 8,000 pediatric patients with type 1 diabetes) and 3,000 controls (1,500 selected from the 1958 BBC and 1,500 UK blood donors). The WTCCC samples have been genotyped using the Affymetrix 500K chip; however, the 1,500 samples from the 1958 BBC controls were also genotyped at the WT Sanger Institute using the Illumina 550K platform. Thus, merging the data from the T1DGC Stage 1 with the WTCCC would yield a collection of 6,000 genotyped T1D cases and 5,500 controls (or 4,000 cases and 4,000 controls genotyped with Illumina technology).

The choice of 4,000 cases from the JDRF/WT British case collection and 2,500 controls from the B1958BC for Stage 1 was based on the immediate availability of both case and control samples (currently at the JDRF/WT Diabetes Inflammation Laboratory (DIL) at Cambridge University) for aliquoting at a single location, homogeneity of study population (Great Britain) with existing and extensive evaluation for population stratification, large sample size within the single population for optimal detection of even minor gene effects and even larger effects for which the genome-wide SNP coverage is not optimal.

The T1DGC Stage 1 data management, clean-up and allele calling, and statistical genetic analyses will be led by Professor David Clayton of the University of Cambridge. Under Professor Clayton's guidance, the Stage 1 data will also be merged with the WTCCC data and analyzed. Results of these analyses will be made available to the T1DGC membership and announced on the T1DGC web site. T1DGC Stage 1 data (4,000 cases and 2,500 controls) will be merged with the WTCCC data on type 1 diabetes to create a merged data set. This merged data set (6,000 cases and 5,500 controls) will be made available by request through the T1DGC Access Committee.

A primary objective of the analysis of the first-stage GWA data is to identify the set of SNPs to be followed in the second stage of the GWA experiment. This objective will be achieved as rapidly and as efficiently as possible by the T1DGC using all the resources, expertise (internal and external) and advice as is available. The primary goal of T1DGC Stage 2 is the selection of the top SNPs from the most highly associated regions obtained from the analyses of the T1DGC Stage 1 merged data sets. Samples will be analyzed using the TaqMan assay.

5. Issues and Outcomes

Obviously one cannot predict how many effects will occur within one linkage peak nor the size of each effect if there are multiple, clustered genes, not the allele frequencies of the etiological variants. The "support intervals" from a linkage analysis are often broad, with hundreds of kb to several Mb covered. The size of the support intervals is dependent upon many factors, but often will include scores of genes. In the genome-wide association scan, similar issues of size of regions occur. For the genome-wide association scans, robustly associated SNPs are identified and the region can be defined by the recombination break points or presence of another robustly associated SNP. While these regions are usually narrower than those of linkage scans, there can still be substantial number of genes (or very few, depending on the region of the genome).

There are multiple approaches to following a genome-wide linkage or association scan. These approaches can be performed in parallel. One approach is to perform dense SNP mapping within the regions of support for linkage or association. These SNPs can be chosen to maximize information content and, with resources from HapMap and 1000 Genomes Project, provide reasonable coverage to SNPs with 0.5% minor allele frequency. An alternative approach to genotyping is DNA sequencing. DNA sequencing can take several approaches. One approach is to sequence the entire region, a process that will capture regulatory, coding (exons) and non-coding (introns, intergenic) regions. A second approach would sequence only the regulatory and coding regions (exons). The choice of sequencing is dependent upon a number of factors, including the coverage of the region, the ability to capture non-repetitive sequence, and interpretation. Once genotyping or sequencing is performed, a number of analytic approaches to discovery of causal genes and causal variants can be conducted. Novel variants from genotyping or sequencing can be classified into groups based upon predicted function (e.g., deleterious or beneficial or neutral) using as series of population genetic and biochemical algorithms. In addition to classical single variant association tests, collapsing (or burden) tests can be performed to count the number of predicted deleterious variants in a gene, thereby providing a gene-based, rather than a variant-based, test statistic.

6. Phenotypic and Laboratory Data

T1DGC collects a minimal amount of phenotypic data, which will be described to characterize the study cohort and linked to the results of genetic analysis. The results of the autoantibody analysis will be included in the phenotypic data set as a binomial variable (i.e., positive or negative) and as a continuous variable with the assay value.

E. ANALYSIS REQUESTS

Requests for central statistical analyses must be approved by the Steering Committee and are limited to either: 1) analyses related to approved publications or presentations; or 2) analyses related to allocating or obtaining resources or study management.

F. DATA DISTRIBUTION

T1DGC data holdings are available for access by T1DGC investigators according to the T1DCG Policy Governing Access to Study Repository Samples and Data. Applications for access are coordinated and reviewed by the Access Committee, which is appointed by the Steering Committee.

T1DGC investigators are permitted to collect data and specimens for separate, independent analysis, in addition to those contributed to the T1DGC. Collection must be governed by local human subject’s guidelines and appropriate consent, but are not subject to T1DGC control or policy.

When the Coordinating Center ceases to function as an analytic resource to T1DGC, it will release a fully documented copy of all study data to the NIDDK Central Repository. Access to T1DGC data will then be governed by the NIDDK Central Repository Access Policies. Confidentiality of individual participants will be maintained with all releases of the data.

1. Format

The T1DGC will develop a standard format for the data sets that it distributes. The distribution of data will be via the T1DGC web site, after successful login.

2. Documentation

The T1DGC Coordinating Center will maintain and distribute documentation of study databases that will include: 1) the source of each variable; 2) a description of the markers and genetic mapping; 3) data quality control practices; 4) formatting; 5) descriptions of any created variables; and 6) contact information.

3. Modified Data Sets

Central study databases will be “published’ as discrete versions according to scheduled negotiated by the Coordinating Center with the Steering Committee. Versions will be updated periodically, as additional data are accrued and as data editing decisions and routines are implemented.

LITERATURE CITED

1. Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* 1996; 58: 1347-1363.
2. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993; 52: 506-516.
3. Martin ER, Monks SA, Warren LL, Kaplan NL. A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am J Hum Genet* 2000; 67: 146-154.
4. Kruglyak L, Lander ES. Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am J Hum Genet* 1995; 57: 439-454.
5. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993; 52: 506-516.
6. Spielman RS, Ewens WJ. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am J Hum Genet* 1998; 62: 450-458.
7. Sham PC, Curtis D. An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Ann Hum Genet* 1995; 59: 323-336.
8. Boehnke M, Langefeld CD. Genetic association mapping based on discordant sib pairs: The discordant alleles test (DAT). Am J Hum Genet 1998;62:950-961.
9. Allison DB. Transmission-disequilibrium tests for quantitative traits. *Am J Hum Genet* 1997; 60: 676-690.
10. Schaid DJ. General score tests for associations of genetic markers with disease using cases and their parents. *Genet Epidemiol* 1996; 13: 423-449.