VII. GENETIC MEASURES

A. BACKGROUND

The etiology of type 1 diabetes is unknown, but it is recognized to be due to both genetic and environmental determinants (1). The genetic basis of type 1 diabetes is complex and likely to be due to genes of both large and small effect. There have been numerous studies investigating genetic susceptibility loci, using both case-control and family study designs. Early studies of disease concordance using twin designs reported higher monozygotic (MZ) than dizygotic (DZ) rates, with MZ rates approaching 50% (2-3). These early studies were likely biased, however, as recruitment of the twins were through advertisement and solicitation, so that affected concordant pairs were more likely to participate than discordant pairs. Population-based twin studies confirmed the increased concordance in MZ pairs, but with the concordance of 30%-40%, and the concordance in DZ pairs only 5%-10% (4-5). Based upon the results of twin studies, susceptibility to type 1 diabetes is determined by genetic risk factors, but less than 50% of the risk is due to the effects of genes. Studies of first and second-degree relatives also support familial aggregation of type 1 diabetes (6-8).

The Consortium will use a well-tested, positional cloning approach to identify genes that influence susceptibility to type 1 diabetes. Positional cloning has been used in studies of many Mendelian and complex disease phenotypes, and is currently considered a standard approach to cloning disease susceptibility genes. By genotyping a comprehensive set of polymorphic markers that are equally-spaced across the genome, and by tracing the familial inheritance of the marker genotypes, it is statistically possible to identify regions of the genome that are disproportionately shared by affected family members, as compared to unaffecteds. These regions are called “linkage regions.” More sophisticated analyses will seek to identify families within the Type 1 Diabetes Genetics Consortium (T1DGC) collection who show linkage to different regions of the genome (heterogeneity), and stratify by highly specific individual genetic markers that have been shown to be associated with risk of developing type 1 diabetes. These specific risk markers include HLA (human leukocyte antigen) loci; single nucleotide polymorphisms (SNPs) clustered around the *CTLA4* (cytotoxic T-lymphocyte-associated protein 4) gene (9) and a SNP in the promoter region of the *INS* (Insulin) gene.

B. GENETIC MARKER TYPING

The Consortium will perform genetic marker typing of study participants to characterize genetic variation at these known markers (landmarks) in the human genome. The genotyping will support three specific analytical modalities:

1. Whole genome scans to identify families that show linkage of disease status to certain regions of the genome, hence focusing the search for susceptibility genes. This analysis seeks to discover broad genomic regions (*e.g.*, 10-20cM in size) where a type 1 diabetes disease susceptibility gene is most likely to be located. With the development of genome-wide SNP marker panels that increase information content

and genomic coverage, the Consortium will continue genome-wide linkage scans in families to identify regions containing type 1 diabetes susceptibility genes. These marker panels contain ~6,000 SNPs placed equidistant in the genome, and are amenable to standard methods of linkage analysis.

2. Stratification of study families and individual participants based on independent markers of genetic risk and assessment of association of type 1 diabetes with these markers. Three regions of the genome have been consistently implicated in risk for type 1 diabetes: HLA (*IDDM1*), *INS* (*IDDM2*) and *CTLA4* (*IDDM7*). Recent evidence suggests that additional genes in the HLA region as well as novel susceptibility loci (*PTPN22*, *IFIH1*) contribute to risk of type 1 diabetes. The Consortium will utilize the existing collection of affected sib pair (ASP) families to further explore the genetic information in the HLA region using ~3,000 SNPs, 66 microsatellite polymorphisms, and classical HLA genotypes in the "MHC Fine Mapping Project." Additional candidate genes that had been previously identified as contributing to risk of type 1 diabetes are being characterized by performing SNP analysis in the "Rapid Response Project."

3. In addition to performing analyses of ASP families (linkage), the Consortium has designed a genome-wide association project in which ~500,000 SNPs are genotyped in 4,000 cases with type 1 diabetes and 2,500 controls. These samples, obtained from the UK and complementary to the Wellcome Trust Case Control Consortium, will provide information that identifies genomic regions containing genes that are associated with risk of type 1 diabetes.

   Genetic data will be obtained from these classes of genetic markers:

   1. Single Nucleotide Polymorphism (SNP) markers used for genome-wide linkage (~6,000 SNPs) or genome-wide association (~500,000 SNPs)
   2. HLA markers
   3. Candidate gene polymorphisms (SNPs) in *CTLA4*, *INS*, *PTPN22*, *IFIH1*, and other genes

The genotypes from these classes will constitute the primary genetic data that will be used for identification of linkage regions in the first phase of positional cloning, by linkage mapping and risk stratification. From the joint information derived through genome-wide linkage and association analyses, two follow-up SNP genotyping experiments were conducted. The first follow-up was based upon the ~50 SNPs identified from the genome-wide association scan meta-analysis in an independent set of cases, controls, and families. The second follow-up was composed of a dense SNP mapping panel developed by a consortium of investigators interested in autoimmunity. This SNP mapping panel was composed of 200,000 SNPs that covered regions of significance for seven primary autoimmune diseases. As many of the 40+ loci identified by the T1DGC were also implicated in other autoimmune diseases, there was a great opportunity to leverage both content of the array and sample size by autoimmune disease consortia. The custom autoimmune mapping array (ImmunoChip) consists of over 180 loci across the

genome. The identity of markers used for the fine-mapping will be provided to the scientific community in 2011.

1. Microsatellite (STRP) Marker Typing

Microsatellite markers (also called STRP, Simple Tandem Repeat Polymorphisms) are non-functional markers that consist of variable number of repeats of (usually) two, three, or four nucleotides and are highly polymorphic. They occur frequently in the human genome as a result of normal DNA mutation processes and are not generally believed to be functionally important. Their utility in routine genome scans comes from their high information content, and because they can be typed by semi-automated, high throughput methods. A variety of chromosome-specific sets of markers optimized for high-efficiency genotyping have been developed.

The Center for Inherited Disease Research (CIDR) is a centralized facility, housed in the Johns Hopkins Bayview Medical Center in Baltimore, which was established to provide genotyping and statistical genetics services for investigators seeking to identify genes that contribute to human disease (http://www.cidr.jhmi.edu). Genotyping at CIDR is performed as a service to the research community. Since the Consortium is funded by NIDDK, the Consortium applied for, and obtained, use of CIDR genotyping services for the study. Consequently, microsatellite genotyping of the existing collection of untyped ASP families (approximately 600, see **Chapter II**, *Objectives*) and all newly recruited ASP families (approximately 2,800; see **Chapter II**, *Objectives*) were genotyped at CIDR, at no cost to the Consortium.

With the advent of SNP genotyping, CIDR transitioned from microsatellite to SNP genotyping for both genome-wide linkage and association scans. The T1DGC had their first collection of families genotyped by both microsatellites (~400 in a panel) and SNPs (~6,000 in a panel); future T1DGC collections were genotyped with the same 6,000 SNP linkage panel. CIDR uses standard methods for genotyping. The CIDR web site provides up-to-date marker panel information (*e.g.*, position, size range, marker type, CEPH sizes and primer pairs).

2. Human Leukocyte Antigen (HLA) Locus Genotyping

There are many HLA genotyping strategies, with differing levels of resolution and number of HLA loci considered. In common with other autoimmune diseases, a person's HLA genotype profile can be a significant genetic risk factor for type 1 diabetes susceptibility. In particular, there is important information for type 1 genetic risk at both the class I (A, B, C) and the class II (DR, DQ, DP) loci. The information from the haplotypes defined by A, B, DR, DQ and the existence of novel associations with DP suggested that complete HLA typing is required to fully deconvolute the risk association. In addition, the increase in resolution of susceptibility by sub-typing DRB1 and DQB1 will prove important in defining interactions between HLA and other regions. For example, on DQB1*0302 haplotypes, DRB1*0401 and *0402 (in Caucasians) and *0405 (in Caucasians, African-Americans and Asians) confer susceptibility, while *0403 (in Caucasians and Chinese) and *0406 (in Japanese) are protective. In contrast, other

DRB1 alleles appear to be risk neutral. Therefore the Consortium will type the HLA loci HLA-A, B, C, DQA1, DQB1, DPA1, DPB1, and DRB1.

To type the loci, the Consortium will use the Dynal reverse dot blot method, implemented on line strips supplied through Roche Molecular Systems (Pleasanton, CA) as developed by Drs. Henry Erlich and Janelle Noble. The following set of seven assays will be utilized for the T1DGC genotyping effort:

1. HLA-A
2. HLA-B
3. HLA-C
4. DQA1/DQB1
5. DPA1/DPB1
6. DRB1gen (also includes *INS* + *CTLA4* SNPs, see below)
7. DRB1sub

With the exception of DRB1, each locus is typed on a single line strip with a single preceding PCR amplification step using one pair of primers. Briefly, after PCR amplification of the variable locus, the denatured product is hybridized to the linear array (line strip) in a Bee Blot machine, and after color development, the strips are scanned and the image processed. From the pattern of amplimer binding to the probes on the line strip, the HLA genotype of the sample can be inferred to four digit resolution. DRB1 is more complex, and uses a low resolution assay DRB1gen followed by one or two subtype assays using the one the six subtyping assays (DRB1sub).

3. Single Nucleotide Polymorphism (SNP) Markers

SNPs are non-functional markers that consist of two alternative forms (alleles). SNPs represent the most frequent form of genetic polymorphism in the human and form the basis of the human HapMap. Their utility in routine genome scans comes from their high information content, large numbers in the genome, and ability to be characterized by semi-automated, high through-put methods. CIDR utilizes panels of SNP markers for genotyping of the existing collection of ASP families using ~6,000 SNPs. CIDR, as well as other public and private entities, use larger panels of SNPs for genotyping cases and controls (typically 317,000 or 500,000 markers).

SNPs can be used to provide important information on genomic regions of risk (MHC Fine Mapping), candidate genes (Rapid Response, including *INS* and *CTLA4*), and genome-wide analysis (genome-wide linkage scan and genome-wide association scan). For validation of high-throughput SNP genotyping results, TaqMan genotyping was also performed by standard methods.

C. PROCESSING OF DNA FOR GENETIC ANALYSIS

Participants that meet study eligibility requirements will be interviewed in their local recruiting study clinic to collect their demographic and clinical data, and each participant will be asked to provide blood samples for genetic analysis, biomarker

analysis, and for cell-line immortalization and repository storage. For each cell line sample, within 24 hours of collection, the clinic will ship barcode labeled tube(s) of heparin-stabilized or CPDA-stabilized whole blood at ambient temperature to the DNA Repository in their network. After receipt of the blood sample(s) at the repository, staff will extract DNA from a separate aliquot of the leukocytes separated from the whole blood, and will store the DNA under appropriate low temperature conditions.

Periodically, each network DNA Repository will ship quantitated samples of DNA suspended in solution to six separate laboratories for genetic analysis (genotyping):

1. Shipment of DNA samples from network DNA Repository to network HLA Laboratory

2. Shipment of DNA samples from network DNA Repository to CIDR at Johns Hopkins, Baltimore MD, USA.

3. Shipment of DNA samples from network DNA Repository to MHC Fine Mapping Laboratory at the Wellcome Trust Sanger Institute, Cambridge, UK.

4. Shipment of DNA samples from network DNA Repository to Rapid Response Laboratory at the Broad Institute, Massachusetts Institute of Technology, Cambridge, MA, USA.

5. Shipment of DNA samples from UK GRID and 1958 British Birth Cohort collections of cases and controls, respectively, to the Genome-Wide Association Study (GWAS) Laboratory (Illumina, Inc., San Diego, CA, USA).

6. Shipment of DNA samples from network DNA Repository to TaqMan Laboratory at the JDRF/Wellcome Trust Diabetes and Inflammation Laboratory, Cambridge, UK.

1. Shipments to HLA Laboratories

Shipments of DNA samples to HLA laboratories will occur intra-network (*i.e.*, a DNA Repository will only ship samples for HLA genotyping to the corresponding network HLA laboratory). These shipments may vary in frequency and will occur when the DNA Repository has processed sufficient samples to fill one or more shipping boxes with samples. Each shipment consists of 92 samples. DNA will be supplied to the HLA genotyping laboratories in screw-capped tubes with bar-coded labels. A minimum of 5 ug of DNA will be provided to HLA laboratories, at a suggested concentration of 20 ng per ul (total volume = 250 ul). Samples from all participants will be genotyped by the HLA laboratories.

2. Shipments to CIDR

On a more infrequent basis, all DNA Repositories in the Consortium will be contacted by the Coordinating Center and requested to ship aliquots of DNA solution to CIDR, Baltimore MD, USA for microsatellite genotyping and/or 6K SNP whole genome

scan. The Coordinating Center will use the central study database to generate lists of T1DGC participant identifiers for individuals who have been newly recruited, and who have not had CIDR perform genotyping using their DNA sample. The DNA Repository will ship 120 µg of DNA per sample, diluted to a constant concentration of 100 ng/µl in sterile water. These samples will be shipped to CIDR in plates supplied by CIDR with CIDR barcode labels. Only ASP family members will be genotyped by CIDR.

3. Shipments to MHC Fine Mapping Laboratory

Once during the study, all DNA Repositories in the Consortium will be contacted by the Coordinating Center and requested to ship aliquots of DNA solution to the MHC Fine Mapping Laboratory, Wellcome Trust Sanger Institute, Cambridge, UK. The Coordinating Center will use a cutoff date to generate lists of T1DGC participant identifiers for individuals who will be included in the MHC Fine Mapping project. The DNA Repository will ship 100 µL of DNA per sample, diluted to a constant concentration of 100 ng/µl (10µg). These samples will be shipped to the MHC Fine Mapping Laboratory in plates. Only ASP family members will be genotyped by the MHC Fine Mapping Laboratory. Note, samples were used in two separate projects (MHC Fine Mapping and Rapid Response 2).

4. Shipments to Rapid Response Laboratory

Once during the study, all DNA Repositories in the Consortium will be contacted by the Coordinating Center and requested to ship aliquots of DNA solution to the Rapid Response Laboratory, Broad Institute, Massachusetts Institute of Technology, Cambridge, MA, USA. The Coordinating Center will use a cutoff date to generate lists of T1DGC participant identifiers for individuals who will be included in the Rapid Response project. The DNA Repository will ship 50 µL of DNA per sample, diluted to a constant concentration of 100 ng/µl (5 µg). These samples will be shipped to the Rapid Response Laboratory in plates. Only ASP family members will be genotyped by the Rapid Response Laboratory. Note, samples were used in two separate projects (Rapid Response and Rapid Response 2).

5. Shipments to Genome Wide Association Laboratory

For the initial Genome-Wide Association Analysis, cases with type 1 diabetes collected as part of the Wellcome Trust/JDRF collection at the University of Cambridge (UK Grid) and controls from the British 1958 Birth Cohort (B1958BC) were shipped to Illumina, Inc., for SNP genotyping. The Consortium and Illumina staff coordinated the selection of samples, the aliquoting of samples, and the shipment of samples to Illumina. Genotypic data were sent to dbGaP (NIH) and to the GWAS Analytic team at the University of Cambridge (Professor David Clayton) for quality control analyses and statistical genetic analyses. Cleaned genetic data were deposited in dbGaP for access by the scientific community and the control data were returned to the B1958BC.

6. Shipments to TaqMan Laboratory

Once during the study, all DNA Repositories in the Consortium will be contacted by the Coordinating Center and requested to ship aliquots of DNA solution to the TaqMan Laboratory, JDRF/ Wellcome Trust Diabetes and Inflammation Laboratory, Cambridge, UK. The Coordinating Center will use a cutoff date to generate lists of T1DGC participant identifiers for individuals who will be included in the TaqMan. The DNA Repository will ship 100 µL of DNA per sample, diluted to a constant concentration of 100 ng/µl (10 µg). These samples will be shipped to the TaqMan Laboratory in plates. Participants recruited as ASP families, trio families, cases and controls were genotyped by the TaqMan laboratory.

D. STUDY DATABASE

As described in *Chapter II, Objectives*, the total collection of T1DGC families were derived from three sources. There are approximately 1,200 families who had been recruited and genotyped prior to the T1DGC; approximately 600 families had been recruited and for whom there was stored DNA, but who had not been genotyped previously; and approximately 2,800 new ASP families as well as new trio families, cases and controls that were recruited under the auspices of this protocol. The study database, housed at the Coordinating Center, will aggregate as much genetic and phenotypic data as possible from the participants recruited prior to the Consortium, and will include all phenotypic and genotype data collected from newly recruited participants.

E. PRIVACY AND CONFIDENTIALITY OF PARTICIPANT GENETIC DATA

The collection and aggregation of participant genetic data into a central study database, and the subsequent release of the data to consortium-internal and -external researchers, requires sensitivity to concerns about the use of the data, and diligence to protect the privacy and confidentiality of study participants. This is especially true when a study has a truly global reach and must deal with cultural and ethnic sensitivities (10).

In addition to a comprehensive multi-layered informed consent process, the Consortium has taken additional specific steps to protect the phenotype and genotype data of individuals. The Coordinating Center has a special role in ensuring that data collection is appropriate, and that privacy and confidentiality are maintained during data collection, analysis, and distribution, and in reporting study results.

1. Anonymity of Study Participants

Local clinics will collect and use personal contact information to recruit and schedule study participants, but no information that could uniquely identify participants will be collected on study paper-based forms. Regional Network Centers will randomly distribute sets of study barcode labels to clinics within their network for the purposes of uniquely labeling study artifacts collected from a study participant (*i.e.*, paper forms and specimen tubes). These label sets contain anonymous barcode labels with preprinted study participant identifiers conforming to a prescribed standard format. Label sets will

be used by a clinic as needed, and the clinic will have sole and private discretion in the assignment of the preprinted anonymous study identification labels to participating families. The local recruitment clinic will retain records to link a study participant identifier to personal contact information. All other sites, including laboratories and the Coordinating Center, will use only the anonymous study identifier and will never have access to participant contact information. The demographic and phenotypic information collected for an individual will not include information that could be used to indirectly infer the identity of a participant, for example local postal code or a more granular geographic identifier. Hence, the data sets will conform to data privacy standards that mandate de-identification. Detailed recruitment or sample tracking reports will only display the anonymous T1DGC participant identifier.

2. Analytical Data Sets

Periodically, phenotype and genetic raw data will be extracted from the participant collection database at the Coordinating Center and used to create secondary databases and data sets for analysis. The analysis will involve preliminary checks of the data quality and integrity followed by statistical analyses of the cleaned data. The clean analytical data sets will be used to both conduct Consortium-level analyses on the data and prepare data sets for distribution to Consortium internal researchers or to authorized external researchers.

To provide an extra level of confidentiality of data, a second family analytical identifier (AID) will be assigned to each study participant, as the first step of the creation of a new analytical data set within the Coordinating Center. AIDs will be assigned to participants using a randomization scheme that randomizes across clinics within a network, and across study recruitment periods. The source network of each participant will still be encoded within the new AID, but the identifier will differ from the study participant identifier during data collection. A single key file will link the data collection study participant identifier to the AID. This file will reside on a secured server within the Coordinating Center, and have file permissions that restrict access to designated data management staff within the Coordinating Center. Analyses on newly collected families and participants within the Coordinating Center will use the AIDs if there is no requirement to access raw collection data and/or collection study participant identifiers. The use of the AIDs as identifiers in distributed data sets is described in the sub-section on distribution of data sets below.

3. Data from Pre-Consortium Families

This protocol does not cover the recruitment and collection of data from type 1 diabetes families for which there is already genome scan/genetic marker data (approximately 1200 families), or for families that have been recruited and for whom there is stored DNA. Aliquots of the stored DNA solution for the pre-existing families will be sent to a variety of genotyping facilities, during the study.

The Coordinating Center will aggregate phenotype and genotype data from both these groups of families as available and will assign unique analytical identifiers (AIDs) to the family participants, in addition to existing anonymous identifiers assigned under

the original recruitment protocol. These AIDs will provide consistency with data from the newly collected families that are recruited under this protocol, as well as providing an extra level of confidentiality.

4. Distribution of Data Sets from Coordinating Center to Researchers

The Coordinating Center will be required to distribute data sets to network PI/researchers and to other researchers who have applied for (and received permission) to analyze certain subsets of data. All data sets will be shipped to researchers with study participants identified using the anonymous, randomized AID. If a data set is distributed to a Contributing Investigator and contains families from the investigator's clinic, upon request, the investigator will be given the translation keys to translate the AIDs to data collection study participant identifiers as assigned within the local network clinic. The Coordinating Center may also directly translate the study identifiers for those families within the data set itself. Contributing Investigators will not be provided with the keys for translation of AIDs for participants recruited outside their clinic. Non-Contributing Investigator internal Consortium researchers and researchers external to the Consortium will only have access to the AIDs and will not be provided with the translation key maps between the AIDs and the study participant identifiers.

5. Inferences from the Genetic Data of a Study Participant

1. Inference of Family Relationships

In the course of analysis of collected family genetic data, it will be necessary to check that the inheritance of genetic markers within a pedigree is consistent with Mendelian segregation. These checks are used to detect errors in data for quality control purposes. In some situations there may be genetic evidence that the self-reported pedigree relationships do not represent the true biological relationships. These discrepancies between the reported and probable biological relationships will often be detectable with high confidence, given the number of independent genetic markers that will be genotyped for a study participant. While family relationships may be adjusted in analytical data sets based on these results, *no genetic-based pedigree relationship testing results will be reported back to any participant, or to the clinics that recruited them.*

2. Inference of Ethnic Ancestry

The multiple independent marker genotypes collected for a participant may permit detailed inference of ethnic ancestry. For example, the predominant ethnic ancestry of an individual may be inferred and also their percentage of ethnicity admixture. The self-reported ethnicity of a participant or family based on answers provided during completion of study forms may not in fact constitute their predominant ancestry. In a similar way to the inference of family relationships, adjustments may be made in analytical data sets, *no genetic-based ethnicity testing results will be reported back to any participant, or to the clinics that recruited them.*

6. Revocation of Informed Consent and Withdrawal from the Study

The Coordinating Center will respect every participant's wishes, as articulated in the informed consent, regarding the use and distribution of his/her samples. A study participant may revoke their informed consent at any time, and request that their

specimens be destroyed and their data purged from study databases and data sets. The Coordinating Center will communicate the appropriate laboratories and clinic to arrange for sample destruction, and will purge both raw collected data and analytical data from its databases and data sets. The Coordinating Center will coordinate removal of the participant data from electronic copies of data sets in possession of internal and external researchers.

7. Security of Data Systems

The Coordinating Center will use secure computer systems and applications to manage study data collection and analysis. The specific computer infrastructure security measures that will be in place are described in **Chapter VIII**, *Data Management*.

LITERATURE CITED

1. Rich SS. Mapping genes in diabetes: a genetic epidemiological perspective. *Diabetes* 1990; 39: 1315-1319.
2. Barnett AH, Eff C, Leslie RDG, Pyke DA. Diabetes in identical twins. A study of 200 pairs. *Diabetologia* 1981; 20: 87-93.
3. Olmos P, A'Hern R, Heaton DA, Millward BA, Risley D, Pyke DA, Leslie RG. The significance of concordance rate for type 1 (insulin-dependent) diabetes in identical twins. *Diabetologia* 1988; 31: 747-750.
4. Kaprio J, Tumiletho J, Koskenvuo M, Romanov K, Renuanen A, Erikson J, Stengaard J, Kesaaniemi YA. Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. *Diabetologia* 1992; 35: 1060-1067.
5. Kumar D, Gemayel NS, Deapen D, Kapadia D, Yamashita PH, Lee M, Dwyer JH, Roy-Burman P, Bray GA, Mack TM. North-American twins with IDDM: genetic, etiological, and clinical significance of disease concordance according to age, zygosity, and the interval after diagnosis in first twin. *Diabetes* 1993; 42: 1351-1363.
6. Degnbol B, Green A. Diabetes mellitus among first- and second-degree relatives of early diabetics. *Ann Hum Genet* 1978; 42: 25-47.
7. Rotter JI. The modes of inheritance of insulin-dependent diabetes mellitus or the genetics of IDDM, no longer a nightmare but still a headache. *Am J Hum Genet* 1981; 33: 835-851.
8. Tillil H, Kobberling J. Age-corrected empirical genetic risk estimates for first-degree relatives of IDDM patients. *Diabetes* 1987; 36: 93-99.
9. Ueda H et al. Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature* 2003 May 29; 423(6939):506-11.
10. Anderlik MR, Rothstein MA. Privacy and confidentiality of genetic information: what rules for the new science? *Annu Rev Genomics Hum Genet* 2001;2: 401-33.