Published in final edited form as:

Diabetes Obes Metab. 2009 February ; 11(0 1): 2–7. doi:10.1111/j.1463-1326.2008.00997.x.

Overview of the MHC fine mapping data

W. M. Brown¹, J. Pierce¹, J. E. Hilner¹, L. H. Perdue¹, K. Lohman¹, L. Li¹, R. B. Venkatesh², S. Hunt², J. C. Mychaleckyj³, P. Deloukas², and the Type 1 Diabetes Genetics Consortium

¹Division of Public Health Sciences, Wake Forest University Health Sciences, Winston-Salem, NC, USA

²Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

³Center for Public Health Genomics, Department of Public Health Sciences, University of Virginia, Charlottesville, VA, USA

Abstract

Aim—The aim of this study was to perform quality control (QC) and initial family-based association analyses on the major histocompatibility complex (MHC) single nucleotide polymorphism (SNP) and microsatellite marker data for the MHC Fine Mapping Workshop through the Type 1 Diabetes Genetics Consortium (T1DGC).

Methods—A random sample of blind duplicates was sent for analysis of QC. DNA samples collected from participants were shipped to the genotyping laboratory from several T1DGC DNA Repository sites. Quality checks including examination of plate-panel yield, marker yield, Hardy–Weinberg equilibrium, mismatch error rate, Mendelian error rate and allele distribution across plates were performed.

Results—Genotypes from 2325 families within nine cohorts were obtained and subjected to QC procedures. The MHC project consisted of three marker panels – two 1536 SNP sets (Illumina Golden Gate platform performed at the Wellcome Trust Sanger Institute, Cambridge, UK) and one 66 microsatellite marker panel (performed at deCODE). In the raw SNP data, the overall concordance rate was 99.1% (± 0.02).

Conclusions—The T1DGC MHC Fine Mapping project resulted in a 2300 family, 9992 genotyped individuals database comprising of two 1536 SNP panels and a 66 microsatellite panel to densely cover the 4 Mb MHC core region for use in statistical genetic analyses.

Keywords

association; HLA; microsatellite; quality control; SNP; type 1 diabetes

Introduction

The Type 1 Diabetes Genetics Consortium (T1DGC, www.t1dgc.org) is an international effort to identify genes that determine an individual's risk for type 1 diabetes (T1D). The creation of a resource base of well-characterized families from multiple ethnic groups is proposed that facilitates the localization and characterization of T1D genes that determine

© 2009 Blackwell Publishing Ltd

Correspondence: Mark Brown, Department of Biostatistics, Division of Public Health Sciences, Wake Forest University Health Sciences, Winston-Salem, NC 27157, USA. **E-mail:** wmbrown@wfubmc.edu

Conflict of interest:

The authors declare that they have no conflict of interests in publishing this article.

disease risk. The strongest and most consistent genetic determinant for T1D is in the human leucocyte antigen (HLA) region. In the T1DGC meta-analysis, the logarithm of the odds (LOD) score for the HLA region was more than 100. As much as 50% of the familial clustering of T1D can be attributed to the HLA region, with an estimated $\lambda_s \sim 3.3$. However, despite the size of the effect, it has not yet been possible to fully define the source of the linkage support in terms of associated polymorphisms. Aside from inadequate sample sizes, this is due, in part, to the unique nature of the 4 Mb major histocompatibility complex (MHC) region. It is very gene-rich (in excess of 200 genes, many with immunological functions) with extensive gene duplication, extreme levels of polymorphism because of the presence of the classical HLA class I and class II genes and regions of very extensive linkage disequilibrium. In addition, the MHC region contains several HLA haplotypes that span over 4 Mb.

The specific objective of the MHC Fine Mapping Project was to establish a dense, highly polymorphic marker set within the 4 Mb core HLA region using T1DGC samples. This effort provided accurate and precise genotyping in DNA from ~2300 families comprising 10 000 individuals. HLA region genotyping was performed on samples from all available family members in the T1DGC. Samples were either genomic DNA purified from either whole blood or immortalized B-cell lines or DNA prepared by whole genome amplification (WGA). In addition, a random sample of blind duplicates (duplicate individuals as well as duplicate families) was sent for analysis of quality control (QC). DNA samples collected from participants were shipped to the genotyping laboratory from several T1DGC DNA Repository sites.

Materials and Methods

Genotyping

The T1DGC MHC Fine Mapping Project consisted of a dense single nucleotide polymorphism (SNP) map and extensive polymorphic microsatellite map to help ensure comprehensive coverage. The SNP map covered the 4 Mb MHC core region (ENSEMBL positions 29.5-33.5 Mb), including the classical HLA loci and other genes. The microsatellite panel was selected to help detect rare haplotypes since some of the T1DGC families originate from non-European ancestral groups.

The MHC Fine Mapping Project data consisted of three marker panels, two 1536 SNP sets [Illumina Golden Gate platform performed at the Wellcome Trust Sanger Institute (WTSI), Cambridge, UK] and one 66 microsatellite marker panel (performed at deCODE genetics). WTSI used Illumina's highly parallel Golden-Gate Genotyping Assay that couples primer extension and oligonucleotide ligation [1]. Two allele-specific oligonucleotides of 40 bases and one locus-specific 60-base primer are required. Fluorescent tags are introduced by universal PCR amplification, and the amplified allele-specific products are captured on a universal fibre optic bead array by hybridization. The array is composed of 96 fibre optic bundles each consisting of 50 k fibres. Beads coated with oligonucleotide probes are attached at the end of each fibre. Each locus-specific primer has a 20-base address sequence matching one of the probes on the universal bead array.

The second 1536 SNP panel included 115 markers that had been attempted in the first oligonucleotide pool assay (OPA) 1536 SNP panel. There were four specific reasons for the redundant markers: varied genotyping scores (20 markers), QC purposes (25 markers), failed SNPs that should have passed (53 markers) and failed SNPs with no alternative proxy ($r^2 > 0.9$; 17 markers). Several markers failed again in the second submission. Of the 115 markers, on average 55.3 (± 8.0) markers were successful in 9765 samples. The median number of successful genotypes was 53, with a minimum of 20 and maximum of 75. In the

raw data, the overall concordance was 99.1% (± 0.02); 226 (2.3%) samples had a concordance rate less than 95%. The majority of samples with high discordance caused significant Mendelian inconsistency errors (MIE) within families, and these samples were removed during the Coordinating Center QC procedures.

Each genotyping service had specific QC procedures. For the Illumina platform at WTSI, genotypes were assigned a confidence score between 0 and 1 (best) by the GENCALL algorithm, which reflects the distance of a sample from the centre of its corresponding cluster (e.g., AA, AB and BB). GENCALL also assigns a score per SNP locus, and a cut-off 0.3 was applied for retaining a SNP (passed). If the total SNP GENCALL score was less <0.3, then the SNP was flagged as failed. If the individual GENCALL score for a single genotype was <0.25, then the genotype was flagged as failed. A 50% quantile GENCALL score for each sample indicated sample quality <0.4, and the entire sample was flagged as failed. Standard Illumina gender markers were included to help determine the participant gender of each sample. When comparing original and QC samples, loci with a call rate of <80% and more than one error per plate were flagged as failed. No filtering on Hardy–Weinberg was performed at this stage. The T1DGC Coordinating Center received all raw (flagged) genotyped samples and the corresponding filtered (passed) genotyped samples and retained all QC duplicate genotype samples.

Samples sent to deCODE genetics for microsatellite typing were arranged in plates comprising 93 project samples and 3 CEPH control samples. Samples were diluted to 15 ng/ μ l in a robot according to their concentration. For each marker, the forward primer is fluorescently labelled. The primer pairs had been extensively tested for optimizing multiplex PCR reactions for cost benefits. Alleles were automatically called using DAC, an allele-calling program developed at deCODE genetics Inc., and the program DECODE GT was used to fractionate called genotypes, according to quality, and to edit when necessary. Statisticians perform quality checks, including examination of plate-panel yield, marker yield, Hardy–Weinberg equilibrium, mismatch error rate, Mendelian error rate and allele distribution across plates.

Samples

Samples were sent for genotyping from 2325 families within nine cohorts. The families selected consisted primarily of nuclear families with an affected sibling pair. For all the genotyping sets, 9992 samples were shipped. The Joslin cohort (whole genome amplified DNA) failed the initial WGA as demonstrated by SNP and microsatellite genotyping. Samples were subjected, therefore, to a new round of WGA, and replacements were provided for genotyping; 13 samples from the original submission had been exhausted and were excluded. Overall, there were 9979 production samples and 339 QC samples sent for genotyping of all the three marker sets. For the first 1536 SNP panel, data for 9841 of the production samples and 333 QC samples were returned. For the second 1536 SNP panel, data for 9881 production samples and 333 QC samples were returned. For the microsatellite panel, data for 9820 original samples and 203 QC samples were returned. Replacement samples were sent for both the OPA1 and the OPA2 SNP panels; no replacement samples were sent for the microsatellite panel. Data for 191 replacement samples for the OPA1 SNP set and 205 replacement samples for the OPA2 SNP set were returned. A summary of the samples set is provided in table 1.

The initial QC procedure of the Coordinating Center reviewed the failed status of the markers and samples based on reports from the genotyping facilities. Using the production and duplicate QC samples, concordance rates were generated between the pairs. This rate was based on both samples having a called genotype for a given SNP. The total of concordant SNPs was divided by the total number of SNPs where both samples had a called

genotyped. 'Missingness' was also examined between the two samples. Samples that were discordant (i.e., concordance rate <98%) were reviewed within families to determine which sample had more Mendelian consistency with the rest of the family members. For samples deemed concordant, the sample that had the greatest number of genotypes overall was preserved for analysis. If a production sample or QC failed genotyping, the sample that passed was preserved for analysis.

Each set of the three marker panels was first reviewed as separate data set. The results of each QC procedure were then compared across all the three analyses to detect similar discordance problems. In each genotyping platform, there were 339 QC samples. In OPA1, three samples failed for both the production and the QC genotyping. In OPA2, four samples failed for both the production and the QC genotyping. For both OPA1 and OPA2, there was one production sample that failed where the QC sample passed and six QC samples that failed where the production sample passed. In both OPA1 and OPA2, all pairs that had genotypes were 98.0% concordant. Where concordance between the production and QC sample was 98.0%, 100 samples in OPA1 and 124 samples in OPA2 had fewer missing genotypes in the QC sample. Where this occurred, the data for the QC sample were substituted for the production sample.

Once the production and the QC sample concordance and comparison was completed, full family structure pedigrees were compiled for three additional QC procedures: checking for Mendelian inconsistency errors, checking for relationship misclassifications and checking for duplicate samples. Families were examined for MIE to help detect relationship misclassifications. Mendelian checks were performed using the software *P_{ED}CHECK* [2]. By summarizing the *P_{ED}CHECK* results, the Coordinating Center obtained a count of MIE within each family. If the total number of inconsistencies was greater than 2% of the total marker set, the family was flagged as problematic and individually reviewed to determine the basis of the errors. From these in-depth reviews, pedigrees can be rearranged, restructured or individuals flagged as completely ungenotyped. To aid in this decision, results from all the three genotyping data sets were reviewed together. As part of this QC procedure, individual loci are independently checked for excess MIE. If a locus has MIE in more than 20% of the families, the locus is deemed to be problematic and the entire locus is completely ungenotyped. Loci with MIE counts between 10% and 20% are reviewed on an individual basis to determine if they were to be classified as problematic.

PREST [3] is a family relationship estimation software package that allows estimation of identical by descent statistics in pair-wise relatives. To aid in determination of family structure problems for the MHC data, we used the *PREST* results from the T1DGC 6K whole genome linkage scan (performed by Center for Inherited Disease Research, Johns Hopkins University, Baltimore, MD, USA). Unfortunately, not all families included in the MHC Fine Mapping project had linkage scan data available during the initial MHC QC procedures. Subsequent to the initial release of the MHC data, we obtained genome-wide linkage scan data and were able to reexamine and correct misclassified family relationships.

To determine duplicate samples, the Coordinating Center reviewed pair-wise comparisons within families and between families. Using *GRAPHICAL RELATIONSHIP REPRESENTATION* (GRR) software [4], we obtained identical by state (IBS) statistics for pair-wise individuals. For those individual pairs with estimated IBS >1.98, we reviewed whether they are within families (i.e., twins or duplicate samples) and between families (i.e., same person belonging to two distinct families, multiple individuals common between two distinct families or duplicate sample between families). In conjunction with the IBS, MIE information was used to determine if a switched/duplicate sample had occurred.

The Coordinating Center took results from all of the standard QC procedures and examined individuals and families across all the genotype data sets. Based on this cumulative data, the Coordinating Center was capable of making a scientifically based decision on whether there have been sample switches, unrelated issues, gender discrepancies or duplicate samples (i.e., within families, across families or twins). After all issues were resolved, data sets were assembled for final MIE checks. Using $P_{EDCHECK}$, families were examined again. Families that continued to exhibit high MIE rates were removed from the data set for another round of QC procedures. The remaining families were deemed to have random MIE and were cleaned accordingly. A family that was deemed problematic and a reasonable solution that was not available were completely removed from the analysis data set. Once families were considered 'clean', the family was included the analysis data file.

Results

The T1DGC MHC Fine Mapping Project resulted in four data set releases. Each release resulted in a modified, updated and more complete version of the previous MHC data set. The main modifications for each release are because of the incorporation of new genotyped samples and resolution of previous removed problematic families. Other T1DGC genotyping projects provided more specific relationship information to help facilitate pedigree structure issues.

The initial data release (2006.09.MHC) included 1818 families from eight cohorts. The genotyping data for the Human Biological Data Interchange (HBDI) cohort (all three panels) and the Joslin cohort (microsatellite) had not been received at the time of this release. For the release, 69 families were removed for further investigation because of high level of Mendelian inconsistencies. In addition, 119 individuals had either a pedigree change (i.e., classified new parent, sample switch or gender reclassification) or were deemed unrelated and were completely ungenotyped.

The second data release (2006.12.MHC) incorporated the HBDI families and the Joslin microsatellite panel, representing 2241 families from all nine cohorts across all the three genotyping panels. For this release, 84 families were removed for further investigation because of a high level of Mendelian inconsistencies and 119 individuals had either a pedigree change (classified new parent, sample switch or gender reclassification) or were deemed unrelated and were completely ungenotyped.

The third data release (2007.02.MHC) included 2321 families from all nine cohorts across all the three genotyping panels. For this release, four families were removed: two families with irreconcilable Mendelian inconsistencies and two families with duplicate genotyping. Prior to this release, the Coordinating Center had received more complete genome-based genotyping from other T1DGC projects. Based on this information, we were able to identify more precisely relationship errors, duplicate samples within pedigrees and duplicate samples between pedigrees. There were 235 individuals who had either a pedigree change (i.e., classified new parent, sample switch or gender reclassification) or were deemed unrelated and were completely ungenotyped.

The fourth data release (2007.11.MHC) included 2300 families (98.5% of the 2325 initially submitted) with at least one individual with genotype data and 9768 individuals (97.9% of the 9992 samples submitted) (table 2) from nine cohorts across the three genotyping panels. The final release incorporated replacement samples for 191 individuals in six cohorts for the OPA1 genotyping panel and 205 individuals in six cohorts for the OPA2 genotyping panel. At this point, the majority of the families in the MHC fine mapping project had linkage scan data from other T1DGC projects to assist in the review and final cleaning. Removal of

duplicate families was performed consistently to match other T1DGC data releases. In this release, 25 families were removed: 2 families with irreconcilable Mendelian inconsistencies and 23 families that matched other families with distinct family identifiers (some across cohorts and some within cohorts). There were 165 individuals who had either a pedigree change (i.e., classified new parent, sample switch or gender reclassification) or were deemed unrelated and completely ungenotyped. There are 18 individuals recognized as monozygotic twins; one of the twins was ungenotyped.

Discussion

The T1DGC MHC Fine Mapping Project resulted in a 2300 family, 9992 genotyped individuals database comprising of two 1536 SNP panels and a 66 microsatellite panel to densely cover the 4 Mb MHC core region. Several challenges arose because of receiving genotyping data from different cohorts at different time stages, necessitating four data releases. One of the unique aspects of the MHC Project was the fact that a majority of the families were included in a 6K linkage scan after receipt of the MHC genotyping. The Coordinating Center was able to use the more informative family data to resolve relationship issues within the MHC data set. The project resulted in the creation of one of the largest data sets representing a resource base of well-characterized families from multiple ethnic groups to be genotyped across the MHC core region.

Acknowledgments

This research utilizes resources provided by the T1DGC, a collaborative clinical study sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases, the National Institute of Allergy and Infectious Diseases, the National Human Genome Research Institute, the National Institute of Child Health and Human Development and the Juvenile Diabetes Research Foundation International and supported by U01 DK062418. This work was supported by the Wellcome Trust.

References

1. Fan JB, Oliphant A, Shen R, et al. Highly parallel SNP genotyping. *Cold Spring Harb Symp Quant Biol.* 2003; 68:69–78. [PubMed: 15338605]
2. O'Connell JR, Weeks DE. PedCheck: a program for identifying marker typing incompatibilities in linkage analysis. *Am J Hum Genet.* 1998; 63:259–266. [PubMed: 9634505]
3. McPeck MS, Sun L. Statistical tests for detection of mis-specified relationships by use of genome-screen data. *Am J Hum Genet.* 2000; 66:1076–1094. [PubMed: 10712219]
4. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. GRR: graphical representation of relationship errors. *Bioinformatics.* 2001; 17:742–743. [PubMed: 11524377]

Table 2
Description of T1DGC MHC Fine Mapping Workshop data following QC procedures

T1DGC: 2007.11.MHC, data characteristics, family summary: post-modification										
Contributor	Families	Families [n (%)] with								
		Parents			Affected full siblings					
		0	1	2	0	1	2	3	4+	Affected half siblings
Asia-Pacific	191	25 (13.1)	52 (27.2)	114 (59.7)			179 (93.7)	9 (4.7)	1 (0.5)	2 (1.0)
Europe	475	48 (10.1)	104 (21.9)	323 (68.0)	2 (0.4)	8 (1.7)	440 (92.6)	22 (4.6)	1 (0.2)	2 (0.4)
North America	334	34 (10.2)	96 (28.7)	204 (61.1)			311 (93.1)	12 (3.6)		3 (0.9)
UK	114	5 (4.4)	24 (21.1)	85 (74.6)			105 (92.1)	3 (2.6)	1 (0.9)	
BDA	418	4 (1.0)	56 (13.4)	358 (85.6)	2 (0.5)	38 (9.1)	376 (90.0)	1 (0.2)		1 (0.2)
Danish	147	14 (9.5)	38 (25.9)	95 (64.6)			130 (88.4)	14 (9.5)		
HBDI	431		10 (2.3)	421 (97.7)			377 (87.5)	40 (9.3)	6 (1.4)	1 (0.2)
Joslin	112	40 (35.7)	22 (19.6)	50 (44.6)			94 (83.9)	6 (5.4)		3 (2.7)
Sardinian	78	2 (2.6)	22 (28.2)	54 (69.2)	1 (1.3)	6 (7.7)	71 (91.0)			
Total	2300	172 (7.5)	424 (18.4)	1704 (74.1)	5 (0.2)	84 (3.7)	2083 (90.6)	107 (4.7)	9 (0.4)	12 (0.5)

MHC, major histocompatibility complex; T1DGC, Type 1 Diabetes Genetics Consortium; BDA, British Diabetes Association; HBDI, Human Biological Data Interchange.